Digital Humanities 2016

From Digitization to Knowledge 2016:

Resources and Methods for Semantic Processing of Digital Works/Texts

Proceedings of the Workshop

Workshop held in conjunction with Digital Humanities 2016

July 11, 2016
Krakow, Poland

Series: Linköping Electronic Conference Proceedings, No. 126
ISSN: 1650-3686
eISSN: 1650-3740
ISBN: 978-91-7685-733-5

Preface

This volume contains the papers presented at the Workshop on Resources and Methods for Semantic Processing of Digital Works/Texts held on July 10, 2016 in Krakow: From Digitization to Knowledge (D2K 2016).

The goal of this workshop was twofold: First, provide a venue for researchers to describe and discuss practical methods and tools used in the construction of semantically annotated text collections, the raw material necessary to build knowledge-rich applications. A second goal was to report on the on-going development of new tools and resources for providing access to the rich information contained in large text collections: Ontologies, framenets, syntactic parsers, semantic parsers, entity linkers, etc. with a specific focus on digital humanities.

The committee decided to accept eight papers, where each submission was reviewed by at least two reviewers. We wish to thank all the authors for the quality of their contributions as well as the members of the program committee for their time and reviews.

The organization of this workshop would not have been possible without the support of the Swedish Research Council and the frame program: *Det digitaliserade samhället*, 'The Digitized Society'. We are also grateful to the EasyChair conference management site and Jonas Wisbrant who helped us produce these proceedings as well as to Linköping University Electronic Press for publishing them.

Pierre Nugues

on behalf of the organizing committee: Lars Borin, University of Gothenburg, Nathalie Fargier, Persée (Université de Lyon, ENS de Lyon, CNRS), Richard Johansson, University of Gothenburg, Marcus Klang, Lund University, Pierre Nugues, Lund University, Nils Reiter, Universität Stuttgart, and Sara Tonelli, Fondazione Bruno Kessler

Program Committee

Lars Borin	Språkbanken, University of Gothenburg
Nathalie Fargier	Persée (Université de Lyon, ENS de Lyon, CNRS)
Richard Johansson	University of Gothenburg
Stefano Menini	Fondazione Bruno Kessler, University of Trento
Pierre Nugues	Department of Computer Science, Lund University
Nils Reiter	Institute of Natural Language Processing, Stuttgart University
Rachele Sprugnoli	Fondazione Bruno Kessler, University of Trento
Sara Tonelli	Fondazione Bruno Kessler

Invited Talk: Exploring Wikipedia as a Text-technological Resource: From Natural Language Processing to Modeling Language Change

Alexander Mehler
in collaboration with Wahed Hemati and Tolga Uslu
Goethe University, Frankfurt

Abstract

Wikipedia and related projects such as Wiktionary are widely used as resources for solving various tasks in Natural Language Processing and Computational Linguistics. In the area of semantics, this relates, for example, to explicit semantic analysis and topic labeling as well as to wikification, ontology mining, merging and enrichment. A central prerequisite of any such endeavor is the accurate and detailed pre-processing of these resources along such tasks as automatic tagging of parts of speech and grammatical categories, disambiguation, dependency parsing, relation extraction and topic modeling. In this talk, I will address these tasks by example of the German Wikipedia. The aim is to show how they scale with the size of Wikipedia subject to its growth over time. This is done as a prerequisite for studying laws of semantic change by analyzing several consecutive stages of Wikipedia thereby giving insights into the time and space complexity of time-related lexical-semantic analyses. This is done from the point of view of complex network theory. We show the dependency of the entropy of lexical networks derived from Wikipedia as a function of time. This is done to get sparse representations of vector embeddings to be used for the various tasks of semantic modeling. A special focus will be on visualizing the output of such analyses with the help of the TextImager. The UIMA-based TextImager automatically extracts a wide range of linguistic information from input texts to derive representational images of these texts.

Biography

Alexander Mehler is professor of Text Technology/Computational Humanities at Goethe University, Frankfurt where he heads the Text Technology Lab (`https://hucompute.org`) as part of the Department of Computer Science and Mathematics. Alexander Mehler has been member of the executive committee of the LOEWE Priority Program Digital Humanities and is currently member of the board of directors of the Centre for the Digital Foundation of Research in the Humanities, Social, and Educational Sciences. He investigates formal, algorithmic models to deepen our understanding of information processing in the humanities. Alexander Mehler examines diachronic, time-dependent as well as synchronic aspects of processing linguistic and non-linguistic, multimodal signs. He works across several disciplines to bridge between computer science on the one hand and corpus-based research in the humanities on the other. To this end, the Text Technology Lab develops information models and algorithms for the analysis of texts, images, and other objects relevant to research in the humanities. As an outcome of this research, resources and tools such as the eHumanities Desktop, the OWLnotator and the TextImager have been or are currently developed in order to provide text-technological sustainability.

Invited Talk: The RetroNews Project

Olivier Hamon
Syllabs, Paris

Abstract

Semantic processing offers new perspectives and adds value to large and heterogeneous corpora. In this talk, I will describe the RetroNews project, led by BNF-Partenariats (a subsidiary of the French National Library) and Immanens (a specialist in Digital Publishing). In the scope of RetroNews, Syllabs has carried out several tasks involving semantic processing with the aim of enriching digitalized old press. Several objectives have been pursued in the creation of a portal for old press (http://www.retronews.fr/), such as promoting corpus navigation and providing relevant bounces, offering precise indexation and advanced faceting, or galvanizing contents by proposing temporal and geographic projections, or theme gathering. In this context, Syllabs tools have analyzed 3 million pages so far and 6 further millions are planned to be done. As I will describe – and my talk will focus on these points – a number of processes are carried out in order to extract themes, named entities and topics, most of them based on existing referentials.

Biography

Olivier Hamon is the research and development manager at Syllabs, in Paris. He studied computer science at the engineering school *Sup Galilée* (Paris XIII, France), and started his career as a research engineer and then project manager at the Evaluations and Language resources Distribution Agency (ELDA), Paris. He holds a Master in artificial intelligence and combinatorial optimization and obtained his PhD on evaluation in natural language processing from the *Laboratoire d'Informatique de Paris-Nord (LIPN – Université Paris XIII)* in 2010, with the design and development of a generic and perennial architecture for evaluation. He has participated in a large number of national and European projects around language technologies (machine translation, terminological extraction, question answering, among others) and their evaluation, as well as in projects dealing with language resources and distribution platforms. In 2014, Olivier Hamon joined Syllabs as a research and development manager, where he supervises research projects on information extraction, classification, natural langage generation, etc., coordinates technical projects and contributes to the general strategy of the company.

Table of Contents

Towards interactive visualization of public discourse in time and space

Lars Borin
Språkbanken • Dept. of Swedish
University of Gothenburg
Sweden
lars.borin@svenska.gu.se

Tomasz Kosiński
Språkbanken • Dept. of Swedish
University of Gothenburg
Sweden
tomasz.kosinski@gu.se

Abstract

We report on a proof-of-concept study where we (1) apply NLP tools for extracting political-discourse topics from a large Swedish Twitter dataset; and (2) design an interactive spatiotemporal visualization application allowing humanities and social-science scholars to explore how the tweet topics vary over space and time.

1 Introduction

Public discourse has been characterized as being "among the most remarkable inventions of the early 19th century" (Nordmark, 2001, 42). It has been repeatedly transformed over its long history; technologies have evolved, new media have appeared, and participation has become increasingly inclusive. The most recent manifestations of public discourse are the various social media that have emerged only over the last decade or so, complementing or perhaps even supplanting traditional print and broadcast media as the main arena of public discourse and opinion formation, involving many more citizens in a much more interactive mode than ever before.

However, there are many questions about public discourse as conducted in social media, questions about the demography and representativity of participation, whether the issues are the same as in traditional media, and whether public opinion formation processes have become fundamentally different as a result.

Social and political scientists are naturally eager to investigate these and other questions, but face the daunting challenge of dealing with the content of big and streaming textual data. Together with researchers in computer science and language technology they are rising to the challenge (e.g. Conover et al., 2011; Sasahara et al., 2013; Preoţiuc-Pietro et al., 2015). There is still ample scope for methodological development in this area, however, and the work presented below is intended as a contribution to digital humanities and social science methodology. We build on an earlier study of political discussion on Twitter, and, reusing the data from that study, we (a) refine the classification of the content of the tweets using state-of-the-art language processing tools (section 2); and (b) develop an interactive visualization application where the spatiotemporal distribution of the tweet topics along with meta information from the analysis can be explored (section 3).

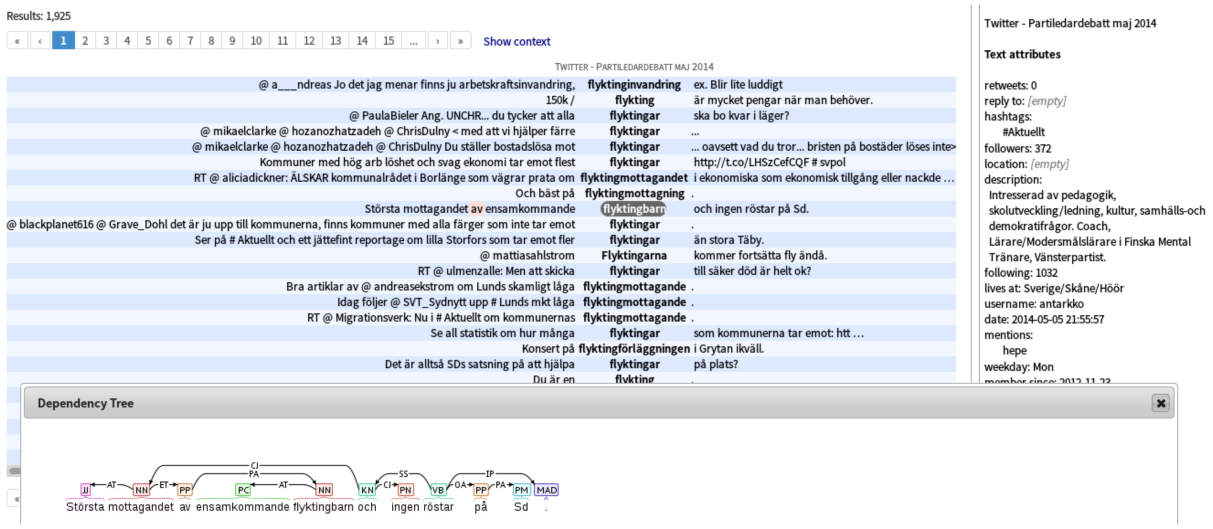
2 Data and research questions

2.1 Studing political debate on Twitter

The data used for the work presented here comes from an earlier study where Swedish tweets were collected from Twitter's public streaming API during a narrow time window around two televised Swedish party leader debates in October 2013 and May 2014, before the national elections in September 2014.

In the earlier study,[1] basic information retrieval techniques were used to classify the tweets into six topics which had been preselected for the debates, and which are considered to reflect two political issue di-

[1] The study referred to is currently under review for a political-science journal, and due to the double-blind nature of this review process, we are unable to reveal the title and authors of this study here.



Results: 1,925

« ‹ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 ... › » Show context

TWITTER - PARTILEDARDEBATT MAJ 2014

@ a___ndreas Jo det jag menar finns ju arbetskraftsinvandring,	**flyktinginvandring**	ex. Blir lite luddigt
150k /	**flykting**	är mycket pengar när man behöver.
@ PaulaBieler Ang. UNCHR... du tycker att alla	**flyktingar**	ska bo kvar i läger?
@ mikaelclarke @ hozanozhatzadeh @ ChrisDulny < med att vi hjälper färre	**flyktingar**	...
@ mikaelclarke @ hozanozhatzadeh @ ChrisDulny Du ställer bostadslösa mot	**flyktingar**	... oavsett vad du tror... bristen på bostäder löses inte>
Kommuner med hög arb löshet och svag ekonomi tar emot flest	**flyktingar**	http://t.co/LHSzCefCQF # svpol
RT @ aliciadickner: ÄLSKAR kommunalrådet i Borlänge som vägrar prata om	**flyktingmottagandet**	i ekonomiska som ekonomisk tillgång eller nackde …
Och bäst på	**flyktingmottagning**	.
Största mottagandet av ensamkommande	**flyktingbarn**	och ingen röstar på Sd.
@ blackplanet616 @ Grave_Dohl det är ju upp till kommunerna, finns kommuner med alla färger som inte tar emot	**flyktingar**	.
Ser på # Aktuellt och ett jättefint reportage om lilla Storfors som tar emot fler	**flyktingar**	än stora Täby.
@ mattiasahlstrom	**Flyktingarna**	kommer fortsätta fly ändå.
RT @ ulmenzalle: Men att skicka	**flyktingar**	till säker död är helt ok?
Bra artiklar av @ andreasekstrom om Lunds skamligt låga	**flyktingmottagande**	.
Idag följer @ SVT_Sydnytt upp # Lunds mkt låga	**flyktingmottagande**	.
RT @ Migrationsverk: Nu i # Aktuellt om kommunernas	**flyktingmottagande**	.
Se all statistik om hur många	**flyktingar**	som kommunerna tar emot: htt …
Konsert på	**flyktingförläggningen**	i Grytan ikväll.
Det är alltså SDs satsning på att hjälpa	**flyktingar**	på plats?
Du är en	**flykting**	.

Twitter - Partiledardebatt maj 2014

Text attributes

retweets: 0
reply to: *[empty]*
hashtags:
#Aktuellt
followers: 372
location: *[empty]*
description:
Intresserad av pedagogik, skolutveckling/ledning, kultur, samhälls-och demokratifrågor. Coach, Lärare/Modersmålslärare i Finska Mental Tränare, Vänsterpartist.
following: 1032
lives at: Sverige/Skåne/Höör
username: antarkko
date: 2014-05-05 21:55:57
mentions:
hepe
weekday: Mon
member since: 2012-11-23

Dependency Tree

×
JJ
NN
PP
PC
NN
KN
PN
VB
PP
PM
MAD
AT
ET
CJ
PA
AT
CJ
SS
OA
IP
PA
Största mottagandet av ensamkommande flyktingbarn och ingen röstar på Sd .

Figure 1: Searching the May 2014 Twitter data using Språkbanken's Korp interface

mensions: *left–right* (topics: *labor market*, *healthcare*, and *education*) and *green/alternative/libertarian–traditional/authoritarian/nationalist* (*GAL–TAN*) (topics: *climate*, *refugees/immigration*, and *crime*).

The tweets were classified into topics using lists of index terms. These lists were incrementally defined by a mixture of manual and automatic methods. Transcriptions of the televised debates and of Swedish parliamentary proceedings formed the basis for the initial, manually constructed, lists. After this, tweets containing at least one word from the initial topic list were merged into a 'topic document' for this particular topic, yielding six multi-tweet topic documents. Additional index terms were identified in these topic documents using the standard tf-idf (term frequency-inverse document frequency) score used for determining index term relevance in information retrieval, and subsequently added to the lists. All list items are text words, and not more abstract linguistic units, such as lemmas or word senses, with the consequence that sometimes several inflected forms of the same lexical item appear in the lists.[2] There are also no multi-word expressions in the lists.

For classification of the tweets, all tweets that contained at least one word from one of the topic lists were considered to discuss the corresponding debate topic. Consequently, tweets could be assigned more than one topic.

The results were presented in numerical form in tables, and additionally in static charts showing preselected subsets of tweet frequencies and topic distributions over time.

Some of the research questions addressed in the earlier study relate to the relative frequency of these topics in the tweets (both in relation to each other and in relation to how much time they were accorded in the televised debates), their timing in relation to that in the debates, and whether *GAL–TAN* issues would be more prominent on Twitter than on television, reflecting a hypothesized difference between professional politicians and social-media users.

2.2 Adding natural language processing

The index word lists used for the classification were kindly made available to us by the authors of the earlier study. The datasets used in their study are available through our research unit – *Språkbanken* (the Swedish Language Bank)[3] at the University of Gothenburg – in the form of annotated corpora, containing user and text metadata, (including location and geographical coordinates) and linguistic annotation of the texts: part of speech, lemma, compound segmentation, word sense(s), and dependency syntax, accessible online through our dedicated web interface for interactive corpus queries, called *Korp*,[4] as well as via

[2] Swedish nouns have 8 different inflected forms, verbs have up to 5 forms, and adjectives have maximally 7 forms.

[3] `https://spraakbanken.gu.se/eng`

[4] `https://spraakbanken.gu.se/korp/`. The corpus import pipeline is available for experimentation through a separate web interface at `https://spraakbanken.gu.se/sparv`.

REST web service APIs and as downloadable datasets in sentence-scrambled form (Borin et al., 2012). See Figure 1, illustrating a corpus search for the lemma *flykting* 'refugee', in all its inflected forms and additionally as part of compounds, e.g., the highlighted word *flyktingbarn* 'refugee children'. The NLP tools forming part of Korp's corpus import pipeline are state of the art, but their performance is unequal and heavily dependent on text type, genre, etc. Adesam et al. (2015) describe ongoing work on building an evaluation dataset which will more faithfully reflect the variety of text types and genres found in our corpus collection, and which consequently will allow us to reach a better estimation of the accuracy of the NLP tools that we use for corpus annotation.

The work presented here is part of a larger effort to design e-science tools for research in the humanities and social sciences (HSS) based on massive amounts of text, richly annotated using state-of-the-art language technologies, providing us with a handle on the content of the texts. There are indications that data visualization and visual analytics have an important role to play here (e.g., Havre et al., 2000; Smith, 2002; Schilit and Kolak, 2008; Chuang et al., 2012; Broadwell and Tangherlini, 2012; Krstajić et al., 2012; Sun et al., 2013), and this aspect is the focus of the work presented here.

Thus, we started out by redesigning the earlier study in this direction. The original word lists – containing text word types, i.e., in many cases several inflected forms of the same lexical entry – were run through an automatic morphological analyzer and the output was manually disambiguated. Unanalyzed words were classified into two groups: (1) simplex words missing from the morphological analyzer's lexicon, in many cases typos or irregular spellings; (2) compounds missing from the lexicon, but having received a compound analysis by the morphological analyzer. The first category was left as-is, while the compounds were (manually) reduced to a common prefix or suffix,[5] e.g., *flyktingorganisation* 'refugee organization', *flyktingproblem* 'refugee problem', *flyktingskatastrof* 'refugee disaster', *flyktingsmuggling* 'refugee smuggling', *flyktingstatus* 'refugee status', are all analyzed as compounds with the prefix `flykting..nn.1` 'refugee n'. Hence, we use only the compound prefix as classification criterion.

This resulted in a considerable reduction in the number of index terms. The average number of words per topic in the original study was 219. The average has now been reduced to 161 index terms (a reduction by 26%), but these of course cover many more text word types.

The topic classification now uses the linguistic annotation layers in addition to the text itself, looking for (a) an exact text-word match (i.e., the only classification criterion used in the original study); (b) a lexical entry match; (c) a compound prefix+compound suffix match; (d) a compound prefix match (e.g., `flykting..nn.1` 'refugee n'); or (e) a compound suffix match, in this order of priority. Note that all but the first capture all the inflected forms of a lexicon word, or a maximum of eight forms for a Swedish noun. Note also that matching for compound parts will result in many more compounds being included than those found in the original lists. As in the original study, a tweet may be assigned to multiple classes.

Our classification results are slightly different from those of the earlier study. Notably, the two most common topics – *labor market* and *education* – switch places. This deserves further study, which however falls outside the scope of this presentation.

It has been frequently observed in the literature that the language of social media deviates in various ways from the written standard language, making the use of off-the-shelf NLP tools problematic. We note here that the word lists used in the earlier study contain predominantly orthographically correct items, and the authors of that study also conducted a small manual check, using lemma searches through a corpus search interface, yielding the same proportions of topics as the automatic classification. However, this procedure only gives us an estimation of the *precision* of the classification, but says nothing about its *recall*, which of course is also dependent on how well the NLP tools work with this text type.

In this connection, we note that the morphological analysis used in the present study is quite reliable, building as it does on a full-sized modern Swedish lexical resource (SALDO; see Borin et al., 2013) with about 140,000 entries, covering on the order of two million inflected forms.[6] However, it does not deal with misspellings or with the various manifestations of creative orthography often encountered in social

[5] Here and below, we use "(compound) prefix" and "(compound) suffix" to refer to the first and second member of binary compounds, respectively, i.e., not in the normal linguistic meaning of the terms "prefix" and "suffix".

[6] `https://spraakbanken.gu.se/eng/resource/saldo`

media, so while the precision is predicted to be high also in our case, the recall is – again – unknown. This is clearly something which deserves further, separate, study.

3 Interactive visualization as a research tool for data exploration

Traditional manual text analysis methods founder when faced with so-called big data, e.g., analyzing thousands of newspapers or millions of blog entries. Human limited cognitive capabilities call for help of machines, which don't get tired or bored, also in this case. Contemporary HSS research already leverages possibilities created by automated tools (Grimmer, 2015) and the computational power available today (Lapponi et al., 2013). But in order to benefit of those fully, the challenges posed by the increasing volumes of data generated and collected everyday and frequently made publicly available on request need to be accounted for and addressed. As already mentioned above, an important emerging technology for dealing with very large amounts of textual data is *visual analytics* (Sun et al., 2013). For a number of practical reasons, in our case, a visual text mining application should preferably be accessible through a web interface.[7] Reviewing existing solutions, the following criteria were taken into account:

(C1) Open-source licensing (to be able to make this work publicly available and open);
(C2) support for real-time, interactive visualization of data amounting to millions or billions of records;
(C3) pixel-oriented technique support (Keim, 2000);
(C4) support for the temporal dimension with real-time, interactive browsing;
(C5) user-defined spatial dimension support;[8] and
(C6) support for browsing individual, non-spatiotemporal dimensions independently.

Our tool of choice, which fullfills criteria (C1-C5), is Nanocubes (Lins et al., 2013), an open-source engine for real-time spatiotemporal data exploration. Criterion (C2) makes it possible to analyse corpora consisting of the amount of source data allowing for representative analysis of textual data sources. Criterion (C3) refers to the relevance of pixel-oriented technique for large spatial datasets visual exploration tools. Criterion (C4) allows for more focused visual search, analysing only a selected time frame at a time, and makes it easier to structure. Criterion (C5) makes it possible to provide non-spatial datasets with a self-designed, simulated spatial domain, supplementing the dataset with a new meaning, integrated with the existing visualization feature, i.e. dimensionality. Criterion (C6) was fullfilled by extending Nanocube's frontend within the presented work.

Visualizing data in two-dimensional space implicitly reduces cognitive load for the user as at least two pieces of information, e.g., latitude and longitude, are presented in the familiar way. All of those features enable real-time sense-making with reproducibility of performed searches, while the user has permanent access to the complete real-world dataset underlying the visualization (Baker et al., 2009).

Using Nanocubes as the data visualization engine, we have established that it is possible to browse information derived from over 20 million Swedish tweets across not only the spatial and temporal dimensions, but also at least 8 other, user-defined dimensions in a highly interactive way. Nanocubes aggregates the data for efficiency and provides no 'way back' to the original data. However, since we believe that this type of visualization will be acceptable to HSS researchers only if they can also at all times inspect the underlying textual data, we have extended Nanocubes with a lookup feature addressing this need. The mechanism behind this feature takes advantage of visual browsing performed by the user with the use of on-screen controls, e.g. buttons, drop-down menus or regions drawn on the top layer of the visualization as well as panning and zooming. Then the user, with each step narrowing down the selected spatial, temporal or categorical dimension, implicitly constructs a corpus query translated into criteria narrowing down the subset of all visualized records. When the user selects the 'dive in' option, he or she is presented with a corpus view of the subset of source material selected based on visual browsing

[7] The reasons for preferring a web interface are not restricted to visual analytics applications, but apply to all kinds of interactive interfaces presenting the results of processing large datasets. A web interface can draw on the generally larger processing and storage capacity of (clusters of) servers, as compared to desktop or laptop computers, so that users can access and process large textual datasets without the need for their own machine to have high-performance or large-storage capabilities. A web interface can further be kept up to date by making changes in one place only, and – quite crucial in many university settings – users will not be dependent on having the administrative privileges required to install client software.

[8] This could be, e.g., a two-dimensional projection of a multidimensional document vector space model.

criteria specified before.

For the proof-of-concept study described here, we used the Swedish Twitter data described above in section 2, providing it with three user-defined dimensions: *Topic* (described above), *Type of match* (which kind of index match was found) and *Strength of evidence* (how many matching words were found). It is evident that this visualization provides added functionality in comparison to the earlier study. Notably, we can explore whether the topic proportions in the tweets are different in different parts of the country (which they seem to be to some extent), selecting moments of interest of the debate. Depending on the resolution of the underlying geolocation data, we can zoom in to even see whether city neighborhoods behave differently with respect to the investigated variables. All the visualizations are available online as a part of bigger work in progress developed to address HSS needs as mentioned before (see Figure 2).

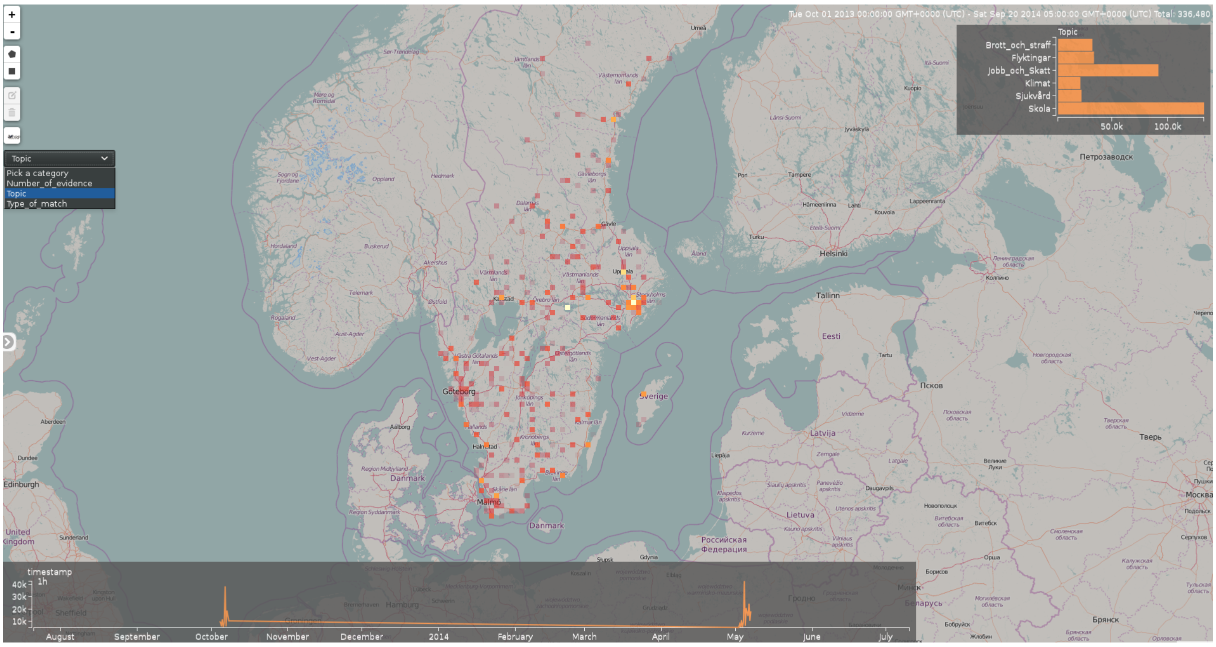


Figure 2: Interactive visualization of the Swedish Twitter debate topics with Nanocubes

There are many issues that remain unaddressed in our small proof-of-concept study. For instance, only about 35% of the tweets could be geolocated. For this we used two kinds of metadata: (1) explicit geographical coordinates, provided in about 17% of all tweets; and (2) matching of words in the "location" metadata against a gazetteer of Swedish place names downloaded from the Swedish postal services, which yielded an additional 18%. Clearly, it would be desirable to do better, perhaps using methods similar to those suggested by Berggren et al. (2015), who geolocate Swedish tweets based on regionally characteristic vocabulary automatically inferred from tweets with explicit location information (mainly proper nouns, but also some dialectal words).

4 Conclusions and future work

We have presented a proof-of-concept interactive spatiotemporal visualization of the results of processing a large Twitter dataset with state-of-the-art NLP tools, enabling more detailed and varied exploration of the research questions of the original study for which the data were collected.

There are several directions in which we intend to continue this work. We think it could be rewarding to enter into a collaboration with the authors of the previous study to explore the usefulness of the kind of spatiotemporal visualization discussed here, as well as investigate the influence on the classification of the NLP tools used. As mentioned above, it is desirable to be able to geolocate more than about a third of the tweets. Also, in order to automate the data pre-processing phase and enable users to visually and interactively analyse the dataset of their choice, the existing visualization engine needs to be integrated with a tool allowing for data preprocessing and formatting, without a limit to the maximal number of records which can be processed.

Other kinds of automated NLP classification will also be added to the datasets as they become available in the corpus import pipeline, e.g., multi-word expressions, word senses, sentiment and argumentation analysis, as well as other methods for topic classification (e.g., LDA or HDP topic modelling), which will help us to throw more light on questions of political opinion formation and expression in social media.

Acknowledgements

This work has been supported by a framework grant (*Towards a knowledge-based culturomics*;[9] contract 2012-5738) as well as funding to Swedish CLARIN (*Swe-Clarin*;[10] contract 2013-2003), both awarded by the Swedish Research Council, and by infrastructure funding granted to Språkbanken by the University of Gothenburg.

References

Yvonne Adesam, Gerlof Bouma, and Richard Johansson. 2015. Defining the Eukalyptus forest – the Koala treebank of Swedish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*, pages 1–9, Vilnius. NEALT.

Jeff Baker, Jim Burkman, and Donald R. Jones. 2009. Using visual representations of data to enhance sensemaking in data exploration tasks. *Journal of the Association of Information Systems*, 10(7):533–559.

Max Berggren, Jussi Karlgren, Robert Östling, and Mikael Parkvall. 2015. Inferring the location of authors from words in their texts. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*, pages 211–218, Vilnius. NEALT.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of språkbanken. In *Proceedings of LREC 2012*, pages 474–478, Istanbul. ELRA.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Peter M. Broadwell and Timothy R. Tangherlini. 2012. TrollFinder: Geo-semantic exploration of a very large corpus of Danish folklore. In *The Third Workshop on Computational Models of Narrative*, pages 50–57, Istanbul. ELRA.

Jason Chuang, Daniel Ramage, Christopher D. Manning, and Jeffrey Heer. 2012. Interpretation and trust: Designing model-driven visualizations for text analysis. In *ACM Human Factors in Computing Systems (CHI)*.

Michael D. Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 89–96, Barcelona. AAAI.

Justin Grimmer. 2015. We are all social scientists now: How big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, 48:80–83, 1.

Susan Havre, Beth Hetzler, and Lucy Nowell. 2000. ThemeRiver: Visualizing theme changes over time. In *IEEE Symposium on Information Visualization, 2000. InfoVis 2000*, pages 115–123, Salt Lake City.

Daniel A. Keim. 2000. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):59–78, January.

Miloš Krstajić, Mohammad Najm-Araghi, Florian Mansmann, and Daniel A. Keim. 2012. Incremental visual text analytics of news story development. In *Conference on Visualization and Data Analysis (VDA '12)*.

Emanuele Lapponi, Erik Velldal, Nikolay Vasov, and Stephan Oepen. 2013. HPC-ready language analysis for human beings. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 447–452, Oslo. NEALT.

Lauro Lins, James T. Klosowski, and Carlos Scheidegger. 2013. Nanocubes for real-time exploration of spatiotemporal datasets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2456–2465, Dec.

[9] `https://spraakbanken.gu.se/eng/culturomics`

[10] `https://sweclarin.se/eng`

Dag Nordmark. 2001. Liberalernas segertåg (1830–1858). In Karl-Erik Gustafsson and Per Rydén, editors, *Den svenska pressens historia. II: Åren då allting hände (1830–1897)*, pages 18–125. Ekerlids förlag, Stockholm.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through Twitter content. In *Proceedings of ACL 2015 (Volume 1: Long Papers)*, pages 1754–1764. ACL.

Kazutoshi Sasahara, Yoshito Hirata, Masashi Toyoda, Masaru Kitsuregawa, and Kazuyuki Aihara. 2013. Quantifying collective attention from Tweet stream. *PLOS ONE*, 8(4):e61823.

Bill N. Schilit and Okan Kolak. 2008. Exploring a digital library through key ideas. In *Proceedings of JCDL'08*, pages 177–186, Pittsburgh. ACM.

David A. Smith. 2002. Detecting and browsing events in unstructured text. In *SIGIR'02*, Tampere. ACM.

Guo-Dao Sun, Ying-Cai Wu, Rong-Hua Liang, and Shi-Xia Liu. 2013. A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. *Journal of Computer Science and Technology*, 28(5):852–867.

The Swedish Culturomics Gigaword Corpus: A One Billion Word Swedish Reference Dataset for NLP

Stian Rødven Eide
Språkbanken
Dept. of Swedish
University of Gothenburg
stian@fripost.org

Nina Tahmasebi
Språkbanken
Dept. of Swedish
University of Gothenburg
nina.tahmasebi@svenska.gu.se

Lars Borin
Språkbanken
Dept. of Swedish
University of Gothenburg
lars.borin@svenska.gu.se

Abstract

In this paper we present a dataset of contemporary Swedish containing one billion words. The dataset consists of a wide range of sources, all annotated using a state-of-the-art corpus annotation pipeline, and is intended to be a static and clearly versioned dataset. This will facilitate reproducibility of experiments across institutions and make it easier to compare NLP algorithms on contemporary Swedish. The dataset contains sentences from 1950 to 2015 and has been carefully designed to feature a good mix of genres balanced over each included decade. The sources include literary, journalistic, academic and legal texts, as well as blogs and web forum entries.

1 Introduction

Having openly available standard datasets for a language is of great benefit for researchers as experiments can be reproduced and algorithms can be compared. A major effort aimed at sharing linguistically annotated Swedish corpora in order to facilitate research in language technology as well as other fields, is ongoing at Språkbanken (the Swedish Language Bank), a language technology research unit and research infrastructure at the University of Gothenburg. However, these corpora typically represent one type of text each, a book, an online forum or governmental reports to give some examples, much like existing datasets for other languages (Sandhaus, 2008; Ferraresi et al., 2008). Following the work of Schäfer and Bildhauer (2012) and others, this paper presents an effort to create a dataset, the Swedish Culturomics Gigaword Word Corpus, where the corpora available for downloading in sentence-scrambled versions from Språkbanken have been sampled to create a large representative contemporary Swedish dataset, from 1950 and onwards. Like the BNC corpus (BNC Consortium, 2007) the aim is to cover different domains and media to offer a balanced dataset. The dataset will be released in clearly indicated static versions to facilitate reproducibility of experiments, comparison of algorithms as well as referencing of data. Each sentence is marked with the year of publication and a genre to help filter the data for specific purposes. It will therefore be possible to use only a portion consisting of, e.g., social media data for a given year.

To assist with common usage scenarios, we will, in addition to the dataset, release code to help extract the text in a desired format; plain text or with different levels of annotation such as part-of-speech tags or multi-word expressions.

2 Contents of the Swedish Culturomics Gigaword Corpus

The dataset contains just over one billion words, sampled from a variety of sources dating from 1950 and onwards. It is designed to be representative of contemporary Swedish, by which we mean texts published from the 1950s until the present day. The last major change in written Swedish was the gradual phasing out of subject–verb agreement. The written standard prescribed the use of distinct singular and plural forms of verbs, even though most spoken varieties had lacked this distinction for centuries. Most authors conformed to the norm until the 1930s, when it became fashionable to switch to using only one form, corresponding to the earlier singular form. Newspapers followed suit in the 1940s, and by 1945, after a recommendation from the newly formed *Nämnden för svensk språkvård* (Swedish Language Cultivation Board), all major newspapers had abandoned the plural form (Pettersson, 1996). By choosing texts published after 1950 we ensure that the grammar of the language is identical to the contemporary form for which Språkbanken's language processing tools are adapted, and hence produce higher-quality annotations.

Each sentence in the dataset is analysed using the corpus import pipeline of Språkbanken's corpus infrastructure *Korp* (Borin et al., 2012). The NLP tools used by the pipeline are state of the art, but their performance is unequal and heavily dependent on text type, genre, etc. Adesam et al. (2014) describe ongoing work on building an evaluation dataset which will more faithfully reflect the variety of text types and genres found in our corpus collection, and which consequently will allow us to reach a better estimation of the accuracy of the NLP tools that we use for corpus annotation. Our dataset becomes a snapshot of all sub-corpora from their creation date and thus constitutes a static dataset which facilitates reproducibility of experiments as well as comparison of algorithms. We trust that this dataset can cover the needs of most NLP research working with contemporary Swedish.

2.1 An overview of our sources

In Table 1 we provide an overview of the various sources included in our dataset, listed with their respective genres and time periods, as well as the numbers of tokens and sentences. We aim to balance the dataset with regards to genre and the date of publication, subject to availability. The fiction genre is particularly small compared to the others because publishing houses have generally been reluctant to release works of fiction for inclusion in corpora. In most of the corpora which can be downloaded from Språkbanken, in order not to infringe copyright the order of the sentences within each corpus have been randomised to limit the possibilities to reconstruct the original text, the freely licensed Wikipedia and official government texts being the notable exceptions.

Source	Genre	Time period	No. of tokens	No. of sentences
Bonniersromaner	Fiction	1976-1981	10,884,795	806,627
Norstedtsromaner	Fiction	1999	2,534,307	194,699
SALT svenska-nederländska	Fiction	1980-1989	1,335,455	96,995
SUC-romaner	Fiction	1990-1999	4,653,784	330,127
Smittskydd	Government	2000-2009	691,716	41,066
Statens offentliga utredningar	Government	1950-1999	50,000,071	2,391,382
Svensk författningssamling	Government	1990-1999	8,335,298	277,030
Svenska partiprogram och valmanifest	Government	2000-2009	821,777	50,684
8 Sidor	News	2000-2009	678,766	59,236
Dagens Nyheter	News	1987	5,122,237	364,226
Göteborgsposten	News	1994-2013	271,239,984	18,935,974
Press 65-98	News	1965-1998	41,177,162	2,891,152
Webbnyheter	News	2001-2013	271,806,921	15,112,300
DiabetologNytt	Science	1996-1999	228,398	14,129
Forskning & Framsteg	Science	1990-1999	744,000	44,538
Humaniora	Science	2010-2015	14,437,043	673,820
Läkartidningen	Science	1996-2005	19,471,910	1,085,785
Samhällsvetenskap	Science	2000-2009	10,873,267	523,102
Svenska Wikipedia	Science	2015	152,333,391	5,972,649
Bloggmix	Social media	1998-2015	35,253,548	2,254,343
Familjeliv	Social media	2000-2015	68,011,169	4,521,566
Flashback	Social media	2000-2015	45,000,152	3,095,212

Table 1: An overview of the sources on which the Swedish Culturomics Dataset is based.

3 Creating the Dataset

For our dataset, we have chosen to keep the XML format inherited from Korp, with some modifications described later in this section. All source texts have been downloaded from Språkbanken's resource pages.[1] The texts have been annotated with the Korp pipeline and the resulting output file is in a simple XML format, distingushing the hierarchical levels *corpus*, *text*, *sentence*, and *w(ord)*.

[1] http://spraakbanken.gu.se/eng/resources

Each word is enclosed in <*w*> tags, which contain syntactical, morphological and semantic annotation from Korp. Much of Korp's word-level analysis is done using SALDO, a lexical-semantic network linking words by their associations, as well as providing information on inflectional morphology and compounding behaviour of lexical items. Although different in several respects, SALDO can be described as a Swedish alternative to WordNet (Borin et al., 2013). The annotations that we extract through the annexed Python code (see section 4) are all derived from SALDO.

While we have kept all of Korp's annotation in our dataset, the code we provide to extract data from it uses either the word itself (for plain text output), the word's *lemma* attribute, the *saldo* attribute (signifying word sense) or the *lex* attribute (the lemgram – a combination of a *lem*ma and *gram*matical information – part-of-speech, inflectional paradigm and compounding behaviour).

In addition to the linguistic annotation, we have added two attributes to each *text* tag; a *year* attribute corresponding to the year of publication and a *genre* attribute chosen from one of the following:

- fiction
- government
- news
- science
- socialmedia

The dataset is structured by decades where each decade consists of several files and each file contains up to one million sentences. This structure, as well as the possibility to filter on genre and year, will allow users to easily choose a subset that suits their purpose, as well as keep the processing requirements low.

4 Using the dataset

The dataset is provided as a series of XML files in UTF-8 encoding, compressed with bzip2 to preserve space and bandwidth. The structure of the XML files is identical to that produced by Korp as described above, with the addition of *year* and *genre* attributes in each *text* tag.

Note that while the order of the sentences is mostly random within each source, we have not performed any additional randomisation when creating the dataset. This means that certain genres may only be found at the beginning or end of the dataset. If randomisation is important for the application, then this must be performed on the dataset. Additionally, some time periods may be dominated by a specific genre, especially the period 1950–1960 for which we only have government-produced text. More recent decades are, however, better balanced.

Distributed with the dataset are two files with Python code[2] that can be used to extract data to a text file also encoded in UTF-8. The code can output any of the following:

- Plain (the original words from the source without any formatting)
- Lemma (each word is replaced by its lemma where Korp has found one)
- Word sense (each word is replaced by its word sense as classified by SALDO)
- Lemgram (each lemgram contains the part-of-speech tag as well as a number signifying the inflectional paradigm)

The output is determined by the `--mode` flag. If this flag is not given by the user, the program will default to plain mode. An overview of the basic usage is provided in Table 2.

For extracting to all outputs except plain, a flag `--mwe` can be used that contracts multi-word expressions (MWEs). In practice, this means that the lemma for the first word in an MWE will contain the whole expression, while the lemmas for the rest of the words are removed from their respective positions in the sentence. The POS is also updated to reflect that it is an MWE. E.g., an adverbial expression such as *en gång till* 'one more time' will receive "abm" as its POS, the final "m" signifying an MWE.

In addition, a flag `--first-only` can be used to only output the first *lemma*, *saldo* or *lex* attribute, respectively, where more than one option is possible. This can be particularly useful for *saldo* output, where the first sense is more likely (Nieto Piña and Johansson, 2016), but less so for *lemma* and *lex*, where any one of the options should be considered equally likely to be correct.

[2] Python 3 is currently required to run the code.

Output	Flag	Attribute	Optional flags
Plain	`--mode`	`plain`	
Lemma	`--mode`	`lemma`	`--mwe --first-only`
Word sense	`--mode`	`saldo`	`--mwe --first-only`
Lemgram	`--mode`	`lex`	`--mwe --first-only`

Table 2: Basic usage of the code to extract data from our dataset.

It is also possible to filter on genre using the `--genre [GENRE]` flag, where `[GENRE]` is one of the above listed genres, written in lowercase. If omitted, the genre flag defaults to all. It is currently not possible to filter on time period without adapting the code, though as the XML files reside in different subfolders from each decade, that should not be necessary in most cases.

4.1 A usage example

A short example that shows how our XML files are annotated is listed here, using the sentence *Hönan lade sina ägg i gräset* 'The hen laid her eggs in the grass'.

```
<sentence id="8f7-8ee">
  <w pos="NN" msd="NN.UTR.SIN.DEF.NOM" lemma="|höna|" lex="|höna..nn.1|" saldo="|höna..1|" prefix="|" suffix
      ="|" ref="1" dephead="2" deprel="SS">Hönan</w>
  <w pos="VB" msd="VB.PRT.AKT" lemma="|lägga|lägga ägg|" lex="|lägga..vb.1|lägga_ägg..vbm.1|" saldo="|lägga
      ..1|lägga..2|lägga..3|lägga_ägg..1|" prefix="|" suffix="|" ref="2" dephead="" deprel="ROOT">lade</w>
  <w pos="PS" msd="PS.UTR+NEU.PLU.DEF" lemma="|sig|" lex="|sig..pn.1|" saldo="|sig..1|" prefix="|" suffix
      ="|" ref="3" dephead="4" deprel="DT">sina</w>
  <w pos="NN" msd="NN.NEU.PLU.IND.NOM" lemma="|ägg|lägga ägg:2|" lex="|ägg..nn.1|lägga_ägg..vbm.1:2|" saldo
      ="|ägg..1|ägg..2|ägg..3|ägg..4|lägga_ägg..1:2|" prefix="|" suffix="|" ref="4" dephead="2" deprel="OO
      ">ägg</w>
  <w pos="PP" msd="PP" lemma="|i|" lex="|i..pp.1|" saldo="|i..2|" prefix="|" suffix="|" ref="5" dephead="2"
      deprel="RA">i</w>
  <w pos="NN" msd="NN.NEU.SIN.DEF.NOM" lemma="|gräs|" lex="|gräs..nn.1|" saldo="|gräs..1|gräs..2|" prefix
      ="|" suffix="|" ref="6" dephead="5" deprel="PA">gräset</w>
  <w pos="MAD" msd="MAD" lemma="|" lex="|" saldo="|" prefix="|" suffix="|" ref="7" dephead="2" deprel="IP
      ">.</w>
</sentence>
```

Using our extraction code, the plain mode would generate the exact sentence as above. The output of the other modes are as follows:

```
$ bw_extract.py --mode lemma
höna lägga sig ägg i gräs .

$ bw_extract.py --mode saldo
höna..1 lägga..1|lägga..2|lägga..3 sig..1 ägg..1|ägg..2|ägg..3|ägg..4 i..2 gräs..1|gräs..2 .

$ bw_extract.py --mode saldo --first-only
höna..1 lägga..1 sig..1 ägg..1 i..2 gräs..1 .

$ bw_extract.py --mode lex
höna..nn.1 lägga..vb.1 sig..pn.1 ägg..nn.1 i..pp.1 gräs..nn.1 .

$ bw_extract.py --mode lemma --mwe
höna lägga ägg sig i gräs .

$ bw_extract.py --mode saldo --mwe
höna..1 lägga_ägg..1 sig..1 i..2 gräs..1 .

$ bw_extract.py --mode lex --mwe
höna..nn.1 lägga_ägg..vbm.1 sig..pn.1 i..pp.1 gräs..nn.1 .
```

4.2 Resource and licence

The dataset described in this article, as well as the annexed code files can be found at `https://spraakbanken.gu.se/eng/resource/gigaword`. They are licensed under the Creative Commons Attribution 4.0 International Licence: `http://creativecommons.org/licenses/by/4.0/`.

4.3 Use cases

A dataset like the gigaword corpus can be highly beneficial not only for language technology in general, as we mentioned in the introduction, but also for Culturomics in particular (Michel et al., 2011), seeing as it makes it possible to track cultural changes as reflected in Swedish texts over time. Examples of use cases for this would be to analyse how attitudes have changed, the emergence of new technologies or to detect shifts in importance for any given topic.

5 Conclusions and future work

In this paper we have described a one billion word corpus of contemporary Swedish, containing sentences from 1950 to 2015. The sentences were chosen to feature a good mix of sources and to be balanced over each decade. The dataset is released with code to help users retain the text in a desired format, with or without annotations. In the future, we plan to release updated versions of the dataset to contain up-to-date texts as well as improved and additional Korp annotations with, for example, word sense disambiguation. We also intend to create embeddings with Word2Vec (Goldberg and Levy, 2014) and make them available together with the corpus.

References

Yvonne Adesam, Lars Borin, Gerlof Bouma, Markus Forsberg, and Richard Johansson. 2014. Koala – korp's linguistic annotations developing an infrastructure for text-based research with high-quality annotations.

BNC Consortium. 2007. The British National Corpus, version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, page 474–478, Istanbul. ELRA.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *In Proceedings of the 4th Web as Corpus Workshop (WAC-4*.

Yoav Goldberg and Omer Levy. 2014. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *CoRR*, abs/1402.3722.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Luis Nieto Piña and Richard Johansson. 2016. Embedding senses for efficient graph-based word sense disambiguation. In *Proceedings of TextGraphs-10*, San Diego, United States.

Gertrud Pettersson. 1996. *Svenska språket under sjuhundra år*. Studentlitteratur, Lund.

E. Sandhaus. 2008. The New York Times Annotated Corpus. *Linguistic Data Consortium, Philadelphia*, 6(12).

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul, Turkey. European Language Resources Association (ELRA).

Extracting Scientists from Wikipedia

Gustaf Harari Ekenstierna
Lund University
Lund, Sweden
dat11gek@student.lu.se

Victor Shu-Ming Lam
Lund University
Lund, Sweden
dat11vla@student.lu.se

Abstract

The Internet is, among other things, a very large and continuously growing source of information and knowledge. This knowledge can be found in the form of text, images, databases, tables, etc. In this article, we describe a system that gathers information from Wikipedia articles and existing data from Wikidata, which is then combined and put in a searchable database. This system is dedicated to making the process of finding scientists both quicker and easier.

1 Introduction

The amount of users on the Internet is growing at a steady rate and so does the amount of information. Statistics from Wikipedia, which describes itself as “the Internet’s largest and most popular general reference work”, show that the amount of text has been growing at a linear rate since 2006 (Wikipedia, 2015). This naturally makes things harder to find, which forces us to come up with new ways of managing the data to make searching, sorting and presentation of it easier. Wikipedia uses a system called Wikidata as its backbone for this job. It is not complete however and is missing a substantial amount of information that is present in the text of the articles.

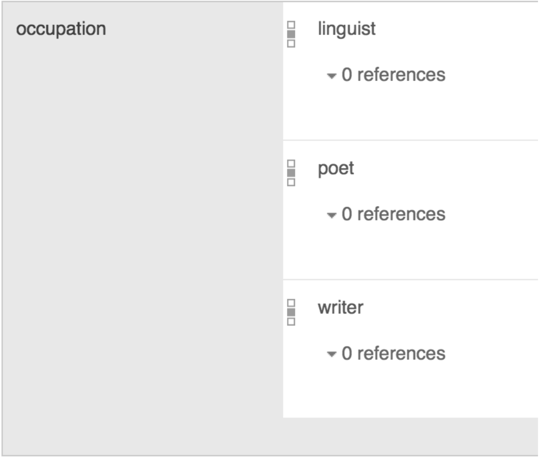
Take Veikko Antero Koskenniemi, a Finnish poet, as an example. In the excerpt below, from his Swedish Wikipedia article, he is described as a literary historian:

> Veikko Antero Koskenniemi, född den 8 juli 1885 i Uleåborg, död den 4 augusti 1962 i Åbo, var en finländsk, finskspråkig författare och litteraturhistoriker.

This fact is confirmed by the article in the English version:

> Veikko Antero Koskenniemi (8 July 1885 – 4 August 1962) was a Finnish poet born in Oulu. In 1921, he took the title of Professor of Literary History in University of Turku, Finland. In 1948, he became a member of the Finnish Academy. He died in Turku.

while his Wikidata profile in Fig. 1 shows a list of three occupations where “literary historian” is absent.



occupation

linguist
0 references

poet
0 references

writer
0 references

Figure 1: Excerpt from Wikidata entry for Veikko Antero Koskenniemi

This issue means that we miss data unless we search both sources. One way of improving the search results is to encourage users to manually register information in the relevant data structures when adding new content, but this is by no means a seamless nor practical solution when considering the large volumes of text. If we want to apply this search to complete digital libraries, the process of taking information from text and putting it into a structured and easily searchable source needs to be automatic.

In this paper, we describe a system, which performs a simple text analysis to extract targeted pieces of information from a given text source. We designed a test procedure to evaluate the contribution of text and early results of our work show that another 70% of information can be found in text in addition to the information contained in Wikidata.

2 The Goal

Extracting people from specific categories is a frequent task: For instance, people from Berlin or Poland, or politicians, etc. Encyclopedias are relevant sources for this task, but as people are often not exhaustively categorized in the text, one needs to read it to find their categories.

If such work could be carried out by programs, we would be able to collect a lot more useful data and statistics with less work. That is what we aim at achieving using a simple text analysis. Instead of reading the articles one by one, we pre-process all the relevant data to create a comprehensive database that can either confirm or deny a person's status as a scientist and point out the references.

For this specific project, we have chosen to extract scientists from Wikipedia as it is a substantial group of people that most of us find interesting. More importantly, Wikipedia has a lot of information provided by the Wikimedia Foundation that we can process. We have created a proof-of-concept by extracting information from text and adding it to an easily searchable database presented on our project website.

To achieve this goal, we had to split our idea into two parts. The first step was to gather all data we needed and to extract only the relevant parts. The second part of our project was to present and make use of our results. We created a web interface where a user can search and explore the gathered information. The website provides a comprehensive preview of our results. However, there are several other ways that our results can be put to use so the website is mostly to demonstrate a case where information extraction can be really useful.

3 Related Work

The idea of extracting information from text is not new. A lot of projects extract information for computational knowledge systems, virtual assistants etc. One of the more known projects is Pantheon 1.0 (Yu et al., 2016), a manually verified dataset of globally famous individuals. They used Freebase(Bollacker et al., 2007) which is a knowledge database very similar to Wikidata and 277 language editions of Wikipedia to produce the dataset. Their first step which was getting the raw data used a process similar to ours by first collecting people entities from Freebase and then mapping those to their Wikipedia articles.

Another large project in language processing with focus on Wikipedia is DBpedia (Lehmann et al., 2014), which aims to extract information from Wikipedia articles into structured data. The resulting data is in many ways similar to Wikidata. DBpedia is limited to Wikipedia data and will only change if the article also changes unlike Freebase which extracts data from multiple sources and combines it with user data to provide a structured data source.

4 The Scope

We are using Wikidata and Wikipedia as information sources. They make a great fit for our project because they are free, closely connected, and contain a lot of information in many different languages. For this paper, we have chosen to limit the scope to content that has a Swedish translation. This decision was not based on any technical limitation. In fact, the project website contains data extracted from both English and Swedish. The tools can easily be adapted to work in any language. But due to time and hardware constraints, we chose to narrow it down to only Swedish to be able to present accurate statistics.

Wikidata is a knowledge database run by the Wikimedia Foundation. It is easily searchable due to its clear structure and therefore makes a good starting point for our project. We used a JSON copy provided by the WikiParq project (Klang and Nugues, 2016). Each entity is represented by an object which holds key-value pairs. The key is called a property, e.g. "occupation" which is an area we are interested in, while the value usually refers to another object, e.g. "scientist". See Fig. 1 for an example.

Wikipedia is an encyclopedia containing about 3 million articles in Swedish and many more if including other languages (Meta, 2016). We used a local copy of the Swedish Wikipedia, also from the WikiParq project. Each article has its own JSON object containing the article's identifier, label, text etc. This makes the data ready to parse since we did not need to strip the text from HTML tags beforehand that otherwise exist in the articles.

Having explained the data sources, it is time to think about what makes a scientist? We decided to base our definition on what the Wikidata says. Each object has a property called "subclass of" which refers to a parent object. With that in mind, we started with the scientist object as the root and simply built a tree from it and all of its subclasses, more on this in Sect. 6.1.

One may object that accurately defining a scientist is not as simple. Our method simply builds on Wikidata and is not more arbitrary than the ontology it proposes.

5 Tools

We used open-source tools in this project. We will go through them below to explain what they are as well as what we use them for and how we use them.

Apache Spark is a great tool for managing very large amounts of data that can be distributed across several machines. We used the Spark SQL Java API to interact with the data from our programs.

To show our results we have deployed a standard Apache web server to provide a simple and accessible user interface, from which users can search the data from our project.

6 The Processing Pipeline

The entire process of extracting scientists is split into three steps, each one having its own program. All three are made by us in Java and are utilizing the Spark SQL Java API.

6.1 Scientific-Professions-Finder

The purpose of this program is to find all scientific professions that exist in the Wikidata. This part is a tree builder which finds all subclasses of a given object. This given object, the root of our tree, is "Q901", the ID of the scientist object. The following query is the first one to be executed after loading the Wikidata into a table.

```
SELECT id, labels FROM professions
LATERAL VIEW explode(property) prop AS p
WHERE p.key = ’P279’
AND p.value = ’Q901’
```

The result of this query contains the ID and a list holding the title in different languages of each profession that is a subclass to scientist. The process is then repeated for each profession it has found until no more subclasses exist. A slightly stylized snippet of the final output can be seen below. It is in English since many professions lack a Swedish title and would not make a good example.

```
,- Q901:scientist
|---, Q169470:physicist
|    |---, Q752129:astrophysicist
|    |   ’--- Q2998308:cosmologist
|    ’--- Q6804564:mechanician
|
```

6.1.1 Wikidata Searcher

The purpose of Wikidata Searcher is to find all people who have one or more scientific profession(s) registered as an occupation, i.e. to find all scientists. It searches through the Wikidata by passing queries to Spark and uses the list of scientific professions given by the previous program. That list may however hold duplicates since it contains a tree structure. Those duplicates need to be removed beforehand, along with the dashes used to indent the rows. This can be done with a simple shell script or similar.

The table of people it is searching in is limited to those who have Wikipedia articles in Swedish, as well as being born between two dates specified by the user. The reason for having these delimiters is to reduce RAM usage and execution time since we are using a single machine to run the programs on. This table will be referred to as the "limited table". Below is an example of a query which returns a list of all people in our limited table who have "scientist" as a registered occupation.

```
SELECT limitedTable.id
FROM limitedTable
LATERAL VIEW explode(property)
prop AS p WHERE p.key = 'P106'
AND p.value = 'Q901'
```

This query is repeated for each scientific profession. Wikidata searcher will create two files once it is done, the first one holds a list of scientists which contains their IDs and scientific professions. The second file holds a list of all the people that were in the limited table, scientists or not. This file will be used by the next and last program. Below is a snippet from the output file containing scientists.

```
urn:wikidata:Q169330
|-Q169470:fysiker
urn:wikidata:Q69571
|-Q593644:kemist
urn:wikidata:Q1702846
|-Q42973:arkitekt
|-Q1792450:konsthistoriker
```

6.1.2 Wikipedia Searcher

The third and final program tries to find scientists in Wikipedia articles by using simple regular expressions. The Wikipedia Searcher loads the file containing a list of people that came from the previous program. It then queries Spark, one time for each person to find his or her article. Each article is analyzed with regular expressions to determine if the person was or is a scientist. The regular expression search is performed once for each profession. An example of what the regular expressions can look like is seen below.

```
"(var|är)\\s.{0,40}kemist"
```

The program, using the regular expressions above, will look for the existence of a substring in each article that matches ”was” or ”is”, followed by a space, up to forty characters of any type and ending with "chemist".

The Wikipedia Searcher will create two files once it has gone through the list of people. The first file contains what has been found and is using the same format as the Wikidata Searcher but with quotes included. Having quotes allows us to quickly judge the accuracy of the results. A short snippet is seen below.

```
urn:wikidata:Q733791
|-Q201788:historiker
"var en finlandssvensk konsthistoriker"
|-Q1792450:konsthistoriker
"var en finlandssvensk konsthistoriker"
urn:wikidata:Q937
|-Q169470:fysiker
```

```
"var en tysk-judisk teoretisk fysiker"
urn:wikidata:Q5726883
|-Q350979:zoolog
"var en finländsk zoolog"
```

The second file contains a list of IDs that belong to people who did not have an article associated with them. This file should be empty since each person must have an article to be included by the Wikidata Searcher. If it is not empty then that means there is a mismatch between the copy of the Wikidata and Wikipedia. This can be caused by a multitude of reasons.

6.2 Data Storage

All the information that the searcher programs have extracted is stored as categorized input/output files for reference. We have also inserted it into tables in a database to have a source optimized for presentation and statistical queries.

The database consists of the following tables:

Professions: A list of all scientific professions with their Wikidata ID, English and Swedish title.

Keywords: A list of all scientific professions with a Swedish title.

Scientists: A list of all the people that have at least one scientific profession, containing their Wikidata ID and Swedish name.

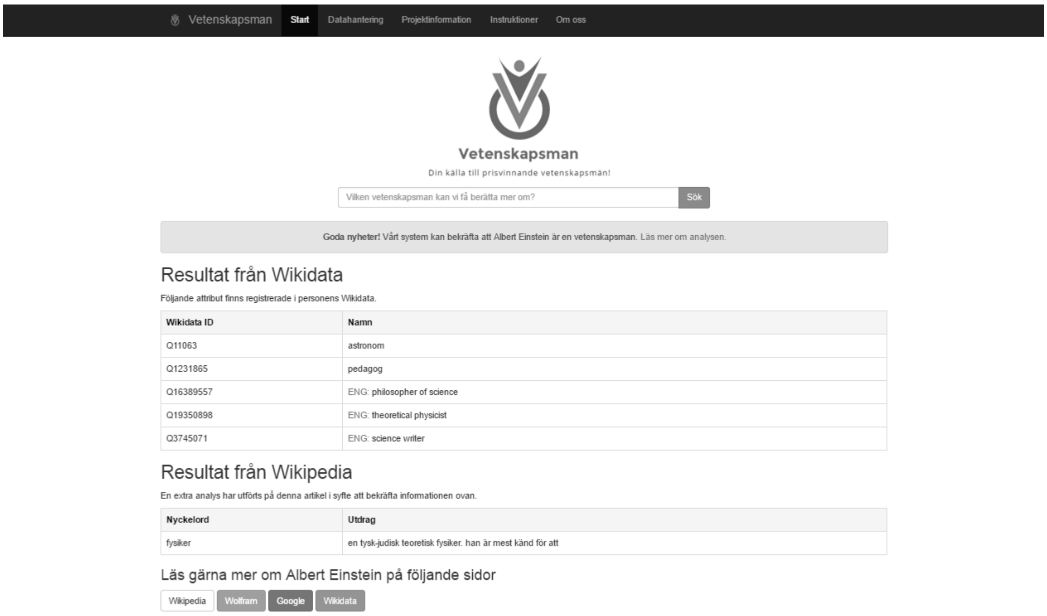
Occupations: A relational table that connects the scientists with their profession.

Quotes: A table with all the extracted quotes from Wikipedia and the Wikidata ID of the scientist described.

6.3 Interface

The resulting data from our work is presented as a search engine, seen in Fig. 2. Users can type the names they have in mind into the search box and the website will not only provide an instant answer but also show you which tags and quotes that can verify it. From there you can navigate to the Wikipedia and Wikidata pages containing those claims. The professions that the searcher programs are looking for are listed on the website as well.

Our website is available at this address `www.vetenskapsman.com` along with instructions on how to use it (in Swedish). “Vetenskapsman” is Swedish for “scientist”.



Goda nyheter! Vårt system kan bekräfta att Albert Einstein är en vetenskapsman. Läs mer om analysen.

Resultat från Wikidata

Följande attribut finns registrerade i personens Wikidata.

Wikidata ID	Namn
Q11063	astronom
Q1231865	pedagog
Q16389557	ENG: philosopher of science
Q19350898	ENG: theoretical physicist
Q3745071	ENG: science writer

Resultat från Wikipedia

En extra analys har utförts på denna artikel i syfte att bekräfta informationen ovan.

Nyckelord	Utdrag
fysiker	en tysk-judisk teoretisk fysiker. han är mest känd för att

Läs gärna mer om Albert Einstein på följande sidor

Figure 2: Web interface of vetenskapsman.com

7 Results

We evaluated the performance of our system using a small test sample of 90 people. They were selected by specifying our delimiters to be people born in March and April 1879 when running the Wikidata Searcher.

7.1 Evaluation

We computed the precision, recall, and F-score to evaluate the accuracy of our programs. They are calculated from knowing the number of true positives, true negatives, false positives, and false negatives. These numbers were acquired by manually checking each person in our test sample. Table 1 shows the results.

	Precision	Recall	F-Score
Wikidata Searcher	100	59.26	74.42
Wikipedia Searcher	95	70.37	80.85

Table 1: Performance figures

Results for the Wikidata Searcher are more dependent on the people maintaining the Wikidata than our program. The cause of this comes from how the program works. It merely fetches information, without any analysis of it, since the Wikidata is already sorted and has good structure. This means that the program will find what is there and will not find what is not there, for better or worse.

The Wikipedia Searcher which uses regular expressions to extract scientists from text had one false positive. The program saves a quote from each of its findings that can help us improve the analysis. The quote that was wrongfully interpreted is as follows:

> var son till ingenjör
> “was son to engineer”

which while conforming to the regular expression does not make the person a scientist.

7.2 Statistics

The combined knowledge from Wikidata and Wikipedia shows that 27 people were scientists and the remaining 63 were not. Neither program managed to find all 27 on its own. However, by adding the results together, we managed to achieve a total recall of 100%. In Fig. 3, we can see results from both programs compared against each other, with an overlap of 8 scientists.

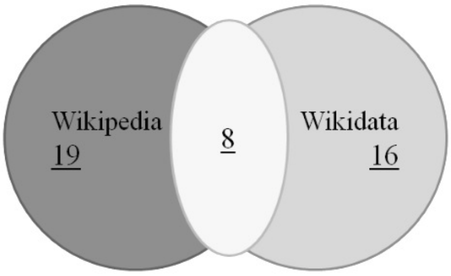


Figure 3: Results from Wikipedia Searcher compared to the Wikidata Searcher

7.3 Future Work

The results look very promising for the Wikipedia Searcher despite its simple regular expression analysis. They may however not be fully representative of the actual accuracy since they are based on a small test sample.

Wikipedia is one of the biggest sources of information and being able to gather data with a real purpose and achieve something useful has been important to us since the start of our project. One of the most interesting and alternative use cases is to use the extracted information to contribute to the Wikidata. The determining factor is that we successfully found many scientists who had their scientific professions missing in the Wikidata. Having built a database of scientists and being able to extract information also helps us build statistics such as how many scientists there are and how that has changed over time. The methods used could also be integrated into the Wikipedia servers to actively create Wikidata by interpreting written articles as well as evaluating the reliability of facts by counting available sources.

8 Conclusions

While the website may be fun to browse and the statistics can be interesting, the user base for such a specific creation is limited but it helped us prove the potential of our idea. With a relatively simple text analysis we managed to find a lot more people who according to our definition are scientists, than what would have been possible by just querying the Wikidata.

If we recall the statistics from our comparison between Wikipedia and Wikidata we can see that out of 27 scientists only 16 were tagged accordingly in the Wikidata. Furthermore, we also know that the Wikipedia Searcher has a precision rate of 95% and the combined precision of both programs is thus 96.43%. What the results told us is that it is possible to increase the number of correctly identified scientists in the Wikidata by almost 70% with a miss rate of about 30% from the Wikipedia Searcher and this is if we exclude the fact that a more advanced analysis could have possibly caught the false negatives that we had with it.

Looking back at the entire project it would be fair to say that the results are satisfying and also better than we could have hoped for considering the relatively simple method that was used and limited time frame. We do see potential for even better results if we were to process more data using a larger collection of keywords and more complex regular expressions and filters. On the other hand, we should have expected good results given the facts that most scientists are clearly described as scientists in their Wikipedia articles and the Wikidata is still new with a lot of work yet to be done. It will be interesting to see how much the Wikidata has improved in a few years from now.

Acknowledgments

We want to thank our instructor Pierre Nugues for his guidance during the project. We also want to thank Marcus Klang for help and tips regarding technical issues and for his work on the WikiParq project. Without the freely licensed data provided by the Wikimedia Foundation and the hard work put into the WikiParq project to format it, our project would not have been possible, we are truly grateful for this. Lastly we want to mention our appreciation to a few developers for various open source frameworks like jQuery, bootstrap and typeahead that proved useful in the development of the website, these include John Resig, Mark Otto, Jacob Thornton, Tim Trueman, Veljko Skarich and Jake Harding.

References

Kurt Bollacker, Patrick Tufts, Tomi Pierce, and Robert Cook. 2007. A platform for scalable, collaborative, structured information integration. In *Intl. Workshop on Information Integration on the Web (IIWeb'07)*.

Marcus Klang and Pierre Nugues. 2016. Wikiparq: A tabulated wikipedia resource using the parquet format. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, Portorož.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Chris Bizer. 2014. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*.

Meta. 2016. *List of Wikipedias — Meta, discussion about Wikimedia projects*. Meta, discussion about Wikimedia projects., https://meta.wikimedia.org/w/index.php?title=List_of_Wikipedias&oldid=15207561. [Online; accessed 12-January-2016].

Wikipedia. 2015. *Wikipedia:Modelling Wikipedia's growth — Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia., https://en.wikipedia.org/w/index.php?title=Wikipedia:Modelling_Wikipedia's_growth&oldid=685257398. [Online; accessed 12-January-2016].

Amy Zhao Yu, Shahar Ronen, Kevin Hu, Tiffany Lu, and César A. Hidalgo. 2016. Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific Data*, 3, 1.

Towards the Automatic Mining of Similes in Literary Texts

Suzanne Mpouli **Jean-Gabriel Ganascia**

Sorbonne Universités, UPMC, CNRS, LIP6 UMR 7606, 4 place Jussieu 75005 Paris
Labex OBVIL, Université Paris-Sorbonne, 1 rue Victor Cousin 75005 Paris
mpouli@acasa.lip6.fr, jean-gabriel.ganascia@lip6.fr

Abstract

Previous studies have shown that not only similes often greatly contribute in establishing the overall tonality of a literary text, but they can also express an author's particular view of the world. This paper presents the architecture of a system geared towards simile mining in literary texts written in English and French as well as some of its early applications.

1 Introduction

Similes can be defined as comparative constructions in which a parallel is drawn between two or more semantically unrelated entities or processes, often through a shared property, so as to produce a mental image in a person's mind. As figures of speech, similes play in essential role in literary texts by making descriptions more vivid and by conveying the right mental image to the readers. While at the syntactic level, they are characterised by specific patterns shared by various languages, semantics enables to distinguish literal comparative statements such as "The pan is as heavy as the pot" from similes like "The pan is as heavy as an elephant's paw". In this respect, studying similes could help to better understand figurative language. Besides, since comparing is a fundamental cognitive activity that relies on individual judgments and associations, similes constitute an interesting basis for studying linguistic creativity.

This paper is organised as follows. Section 2 gives an overview of the different modules of our system. Section 3 summarises the results of preliminary experiments realised on a corpus of French and British novels. Finally, we conclude with perspectives for future work.

2 Overview of the Simile Annotation System

For similes to be annotated in a given text, they first need to be detected. The prototypical simile "The pan is heavy like an elephant's paw", can be represented as "A Ω *y* X B", where A is any type of noun phrase, Ω is a verb, *y* is an adjective, X is a marker of comparison and C is a noun-headed noun phrase. In scholarly works discussing similes, A is commonly referred to as the tenor, Ω or *y* as the ground and B as the vehicle (Figure 1) (Fishelov, 1993).

The pan	is heavy	like	an elephant's paw.
tenor	*ground*	*marker*	*vehicle*

Figure 1. Constituents of the simile "The pot is heavy like an elephant's paw".

In practice, our system is made up of three main modules: a syntactic module, a semantic module, and an annotation module.

2.1 The Syntactic Module

This module is concerned with several preprocessing tasks (tokenisation, lemmatisation, sentence detection, part-of-speech tagging, syntactic parsing), the selection of simile candidates and the identification of each of their components. Since similes can take various forms, a system which seeks to accurately detect similes in texts should be flexible enough to adapt to various simile structures and markers as well as to take into consideration the ambiguity inherent to some simile constructions. In addition to the grammatical markers of comparison of both languages, were also considered other types of markers implying comparisons such as verbal, prepositional and adjectival phrases (see Table 1).

	Comparatives	**Verbal phrases**	**Adjectival phrases**	**Prepositional phrases**
English	*like, unlike, as, as...as, more...than, less...than*	*resemble, remind, compare, seem, verb + less than, verb + more than, be/become… kind/sort/type of*	*similar to, akin to, identical to, analogous to, comparable to, compared to reminiscent of, noun+-like , noun+colour*	
French	*comme, ainsi que, de même que, autant que, plus...que, tel que, moins...que aussi...que*	*ressembler à, sembler, , rappeler, faire l'effet de, faire penser à, faire songer à, donner l'impression de, avoir l'air de, verb + plus que, verb + moins que, être/devenir…espèce/type/genre/sorte de*	*identique à, tel, semblable à, pareil à, similaire à, analogue à, égal à, comparable à*	*à l'image de, à l'instar de, à la manière de à l'égal de, à la manière de, à la façon de*

Table 1. Similes markers for English and French

The selection of simile candidates starts with finding one of the markers in a sentence. Then, the other elements of the potential simile are identified according to their most probable relationship to the last known component, their nature, and their grammatical function (see Figure 2). By default, if various terms can be tagged as tenors or grounds, all plausible elements are extracted. A filtering, which uses manual rules and semantic information, takes place at a later stage to keep only the most pertinent labelled elements.

Vehicle	**Marker**	**Ground**	**Tenor**
Non-subject noun head of the noun phrase following the marker	All except -like and -colour	Verb (+adjective) Non-predicative adjective	Verb subject or direct object Noun phrase modified by the adjective Noun phrase 1
First noun of the compound adjective containing the marker	-like and -colour	Verb	Verb subject or direct object Noun phrase 3

Figure 2. Possible syntactic scenarios.

From these different scenarios, it can be seen that grammatical roles are crucial to the extraction of simile components. It is, therefore, not surprising that dependency parsing had been previously used for this purpose (Niculae & Danescu-Niculescu-Mizil, 2014). We choose, however, to rely on syntactic chunking because our approach is mainly phrase-based and syntactic chunking tends to be more reliable when it comes to capturing close relationships as it is often the case with the ground and the marker. As far as the filtering of wrongly extracted components is concerned, we experimented with agreement rules, lists of transitive verbs and the Sketch Engine (Kilgariff et al., 2004) to extract corpus information on the use of the vehicle as the subject or the direct object of the found verb). Our preliminary tests on a corpus of French prose poems show that our approach (Recall: 66.5%, Precision: 66.3%) performs slightly better than the one based on dependency parsing (Recall: 62.4%; Precision: 64.2%) but captures too much candidates for tenors and has problems with long dependencies. In this respect, for better coverage, we are currently working on blending the two approaches.

2.2 The Semantic Module

Once all the simile components have been found, the next step is to determine whether they express a simile or a literal statement. At least one of the following conditions must be fulfilled for a comparative construction to be considered a simile:
- the ground + vehicle combination is recorded in a precompiled list of idiomatic similes;
- the ground expresses common conceptions about the vehicle, for example, ‘calm’ and ‘lake’;
- the vehicle is part of an extended noun phrase in a comparison of equality;
- the vehicle and the tenor are nouns belonging either to distinct semantic categories or to different subcategories of a broad semantic category (e.g ‘penguins’ and ‘wolves’).

Since accepted ideas about a particular word are connected to its usage, they are embedded in language. Consequently, we put together various French and English machine-readable dictionaries[1] to automatically retrieve specific linguistic pairs: nominal subject-verb, verb-nominal direct object, nominal subject-predicative adjective, adjective-noun. In addition, coordinated nouns, verbs, and adjectives are clustered together as synonyms.

Example: Adjectives frequently associated with ‘biscuit’: flavoured, crisp, rectangular, hard, crescent-shaped, flat, thin, crushed, dry, individual, burnt, unleavened, soft, oblong, German, small.

2.3 The Annotation Module

Although there exist some annotated corpora of similes, none has been devoted to the description of figures of speech for literary studies. Furthermore, despite briefly touching the question of metaphor annotation, the TEI guidelines do not provide a definitive framework, leaving the choice to the encoder.

We distinguish two levels of annotations:
- descriptive annotations which indicate the boundaries of the different elements of the simile, and state for the tenor, the ground, and the vehicle, both the whole phrase they are part of and its head;
- and analytical annotations which provide information about the semantic category of the tenor and/or the vehicle, the idiomaticity of the simile, its frequency in literary texts, and the fixedness of the couple tenor-vehicle or of the triplet tenor-ground-vehicle.

The semantic categories will be derived from coupling each common noun with the clusters of coordinated nouns and other lexicographical information obtained in the previous step. In contrast, for the remaining annotations, a series of experiments were designed to produce a basis to sustain future simile annotations.

Example: The pan is heavy like an elephant’s paw.

```
<creative>
<tenor marker_id="4"> The <head lemma="pan" postag="NN" category="object"> pan
</head></tenor> is <ground marker_id="4"><head lemma="heavy" postag="JJ "> heavy
</head><marker lemma="like" marker_id="4" syntax="null"> like </marker> <vehicle
marker_id="4" >an elephant’s<head lemma="paw" postag="NN" category="body part"> paw
</head></vehicle>.
</creative>
```

3 Examples of Experiments

For the purpose of this project, a reference corpus was put together. In total, 746 French novels and 1190 British novels written between the 1810s and the 1940s were downloaded using online digital libraries such as the Project Gutenberg[2] and the Bibliothèque électronique du Québec.[3]

The first experiment (Mpouli & Ganascia, 2016a) is focused on finding frozen similes in the reference corpus and on determining which ones are used as literary clichés. Two main patterns of frozen similes

[1] French: 8th and 9th editions of Le Dictionnaire de l’Académie française, Littré, Wiktionary; English: GCIDE, Wordnet, Wiktionary

[2] www.gutenberg.org

[3] beq.ebooksgratuits.com

were predefined: adjectival ground + simile marker + nominal vehicle (e.g. *happy as a lark)* and verbal ground + simile marker + nominal vehicle (e.g. *sleep like a top*). The generated results suggest that frozen similes are not so frequent in literary texts, which tends to sustain the idea that creativity plays a central role in literature. Interestingly, English and French share the same most frequent simile “pale as death” or “pâle comme la mort” with 152 and 182 occurrences respectively. In addition, to give new to frozen similes, novelists are very fond of replacing the verbal or the adjectival ground by a synonym or the canonical vehicle by a related noun or an extended noun phrase.

The second and last experiment (Mpouli & Ganascia, 2016b) studies noun+colour term (CT) similes of the type “storm-green sky” in order to investigate if their use of colours correlates the Berlin and Kay’s hypothesis (1969) and how they differ from other similes. From the obtained results, there is no doubt that both noun+CT similes and fully-fledged similes function differently: while traditional similes are strongly governed by collocations and can be used figuratively more easily, noun+CT similes typically describe background elements. This dichotomy could perhaps explain the difference in their use of colours, confirming the idea that colours should not be taken in abstraction, but must be studied in a specific context. Moreover, it is possible to notice an increase in the number of noun+CT similes between the 19th and the 20th century, which suggests that noun+CT similes actively participate in shaping depictions in Modernist novels.

4 Conclusion

Simile annotation is an interesting natural language processing problem that requires not only syntactic but also lexical and semantic information. Apart from improving the identification of simile components, our system also proposes solutions to analyse different types of simile structures. The next phase, besides completing the implementation of the last modules, concerns the evaluation of the system. In this respect, in order to create a gold standard, a platform has been developed to collect annotations of similes in preselected prose poems.[4]

Acknowledgements

This work was supported by French state funds managed the ANR within the Investissements d'Avenir programme under the reference ANR-11-IDEX-0004-02.

References

Berlin, B, & Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution*. Berkeley and Los Angeles: University of California Press.

Fishelov, D. (1993). Poetic and Non-Poetic Simile: Structure, Semantics, Rhetoric. *Poetics Today*, 14(1), 1-23.

Kilgariff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004) The Sketch Engine. *Proceedings of EURALEX*, 105-115.

Mpouli, S., & Ganascia, J.-G. (2016a). "Pale as death" or "pâle comme la mort": Frozen similes used as literary clichés. *EUROPHRAS 2015: Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, 179-187.

Mpouli, S., & Ganascia, J.-G. (2016b; forthcoming). Another Face of Literary Similes: A Study of Noun+Colour Term Adjectives. Selected Papers of the Corpus Linguistics in France Conference.

Niculae, V., & Danescu-Niculescu-Mizil, C. (2014). Brighter than Gold: Figurative Language in User Generated Comparisons. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2008-2018.

[4] French version: dissimilitudes.lip6.fr:8180.

An Overview of Knowledge Extraction Projects in the NLP group at Lund University

Pierre Nugues
Department of Computer Science
Lund University, Lund, Sweden
Pierre.Nugues@ce.lth.se

Abstract

In this paper, I describe systems and prototypes we created in the natural language processing group at Lund to extract structured knowledge from text. Starting from syntactic and semantic parsing components, we developed applications that can handle large corpora, typically complete Wikipedia versions consisting of millions of documents and process text to identify entities and the relations between them. I describe the overall goals of our projects, the data structure we designed to handle the documents, as well as three applications to extract knowledge from text.

1 Question-Answering Systems for Swedish and Other Languages

The multiple digitization initiatives make larger and larger quantities of text everyday more accessible. Within one or two decades, we can imagine that most of what has been written and made public in a printed form will be available in a machine-readable format to anyone with an internet connection. Bill Gates prediction of *information at your fingertips*, made in November 1990, will have come to a reality.

Large parts of the human knowledge are crystalized in text and given its accessibility in a digital form, machines can automatically extract it and process it. The recent success of question-answering systems like IBM Watson (Ferrucci, 2012) that answer questions better than human beings in quiz contests is a proof of it. Text is in fact the raw material of question-answering systems and Watson that builds on the whole Wikipedia collection and documents retrieved from the web, is a dramatic achievement of automatic knowledge extraction.

Most efforts in the development of knowledge extraction systems focus on English. See the evaluations of the Text Retrieval Conferences (TREC)[1], for instance. Such a focus may eventually overlook many sources in other languages and impoverish human knowledge in general. From the beginning, starting from Swedish, our group tried to develop multilingual systems.

In its simplest form, the structure of a question–answering system consists of three components (Fig 1, simplified from Watson):

1. A question processing module that analyzes the question, identifies the entities, and predicts the answer type, a person, location, etc.;

2. A passage retrieval module that builds on knowledge sources, large text repositories, and indexes them. Given a question, this module produces a list of passages that hopefully contain the answer;

3. Finally, an answer extraction module that extracts answers from the passages and ranks them.

We conducted pilot implementations of a question-answering system for Swedish to see how this architecture could generalize. Using a corpus of questions inspired by the Swedish television quiz show Kvitt eller Dubbelt – Tiotusenkronorsfrågan (Thorsvad and Thorsvad 2005; Kvitt eller Dubbelt 2013). (Thorsvad and Thorsvad, 2005; Kvi, 2013), we evaluated the coverage of the Swedish version of Wikipedia as knowledge source (Pyykkö et al., 2014). We split Wikipedia into paragraphs; we indexed them with the Lucene tool; and given a question, we ranked the paragraphs with the $TF.IDF$ measure.

[1] http://trec.nist.gov

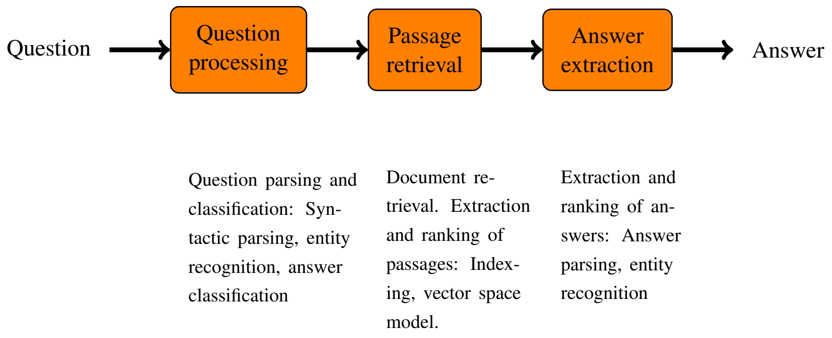

Question
Question processing
Passage retrieval
Answer extraction
Answer
Question parsing and classification: Syntactic parsing, entity recognition, answer classification
Document retrieval. Extraction and ranking of passages: Indexing, vector space model.
Extraction and ranking of answers: Answer parsing, entity recognition

Figure 1: Overall architecture of a question–answering system, simplified from Ferrucci (2012)

We found that about 90% of the answers are in one or more passages of Wikipedia and about 70% in the 200 first passages returned by a $TF.IDF$ ranking. In a second experiment, using the parts of speech of the words in the passages, an answer type predictor, and the category of the proper nouns, we could extract answers with a median rank of the correct answer of 21; we could improve this rank to 10 with a reranker (Grundström and Nugues, 2014). These results showed that the Swedish Wikipedia, although much smaller than the English counterpart, was a viable and valuable knowledge source for question answering systems. This also hinted at a probable replicability of the results to other languages.

2 Propositions Databases

In addition to passages, we can extract structured propositions, consisting of predicates and arguments, from text. For example, from the Wikipedia excerpt:

> Shakespeare was born and brought up in Stratford-upon-Avon, Warwickshire,

we can derive the two predicate–argument structures:

```
born(Shakespeare, Stratford-upon-Avon\, Warwickshire)
brought_up(Shakespeare, Stratford-upon-Avon\, Warwickshire).
```

that convert this piece of text into database facts. Repositories of such facts extracted from large corpora usually improve the performance of question–answering systems – 2.4% in the case of Watson –.

Exner and Nugues (2012) applied a semantic role labeler (Johansson and Nugues, 2008; Björkelund et al., 2010) to the whole English Wikipedia to automatically derive a set of predicate–argument structures. This eventually resulted in more than 260 million propositions (Exner and Nugues, 2014). Figure 2 shows the web interface to the Athena database, where a user can enter a predicate and up to four arguments. In Fig. 2, the user has entered the predicate *kill* and the direct object *bacteria* corresponding to the question *What kills bacteria?* and the system retrieves all the propositions matching the predicate:

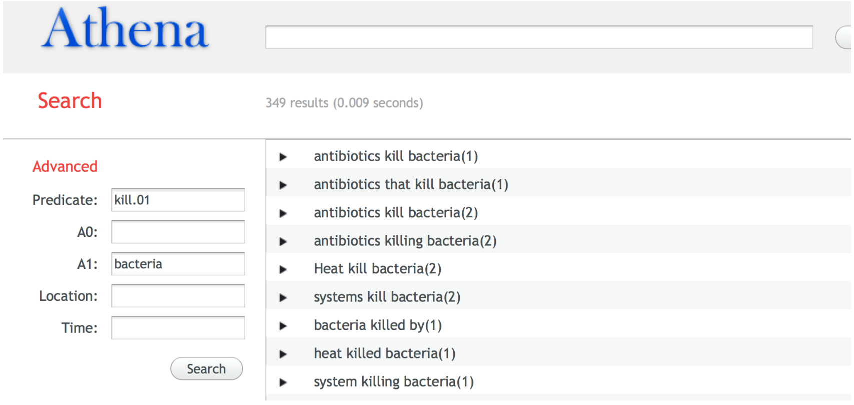
```
kill(X, bacteria)
```

where `X` is a variable, (A0 in Fig. 2), yielding `X = antibiotics`, `X = heat`, `X = systems`, `X = acids`, etc.

3 Multilingual Propositions Databases

While semantic role labelers are generic tools to extract predicate–argument structures, they are language-dependent and require to be trained on large manually-annotated resources. There are no such large annotated corpora for Swedish and many other languages, including French as of the date this paper is written. We developed a tool to project propositions across languages. Fillmore (1976) gives an argument on the crosslingual nature of predicate–argument structures (frames):

> A particularly important notion [...] that goes by such names as “frames”, [...]. Briefly, the idea is that people have in memory an inventory of schemata for structuring, classifying, and



Athena

Search

349 results (0.009 seconds)

Advanced

Predicate: kill.01

A0:

A1: bacteria

Location:

Time:

Search

- ▶ antibiotics kill bacteria(1)
- ▶ antibiotics that kill bacteria(1)
- ▶ antibiotics kill bacteria(2)
- ▶ antibiotics killing bacteria(2)
- ▶ Heat kill bacteria(2)
- ▶ systems kill bacteria(2)
- ▶ bacteria killed by(1)
- ▶ heat killed bacteria(1)
- ▶ system killing bacteria(1)

Figure 2: A screenshot of the Athena system (Exner and Nugues, 2012)

> interpreting experiences, [...] The concept of frame does not depend on language, but as applied to language processing the notion figures in the following way. Particular words or speech formulas, or particular grammatical choices, are associated in memory with particular frames

To extract and abstract such frames across languages, we applied the following ideas:

1. Ground the frames (schematas) in reality through their actors (arguments), independently from the language;

2. Use real world entities, such as Aristotle or *the Organon*, to identify the actors more easily;

3. Find the predicates or verbal nodes connecting these actors.

As an example, the sentence *Aristotle wrote the Organon* has 32 occurrences in Google books[2], while the equivalent French sentence *Aristote a écrit l'Organon* has 3. Recognizing the two named entities, Aristotle and Organon, in propositions, if frequent enough, will probably involve the same relation: write/écrire (Fig. 3) with its two core arguments in Framenet: *author* and *text* (Ruppenhofer et al., 2005). It will be then possible to derive an annotated corpus of relations in French.

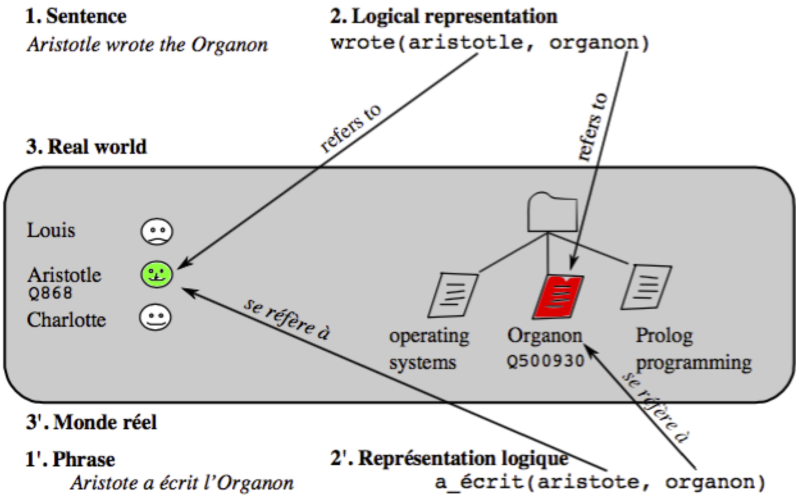
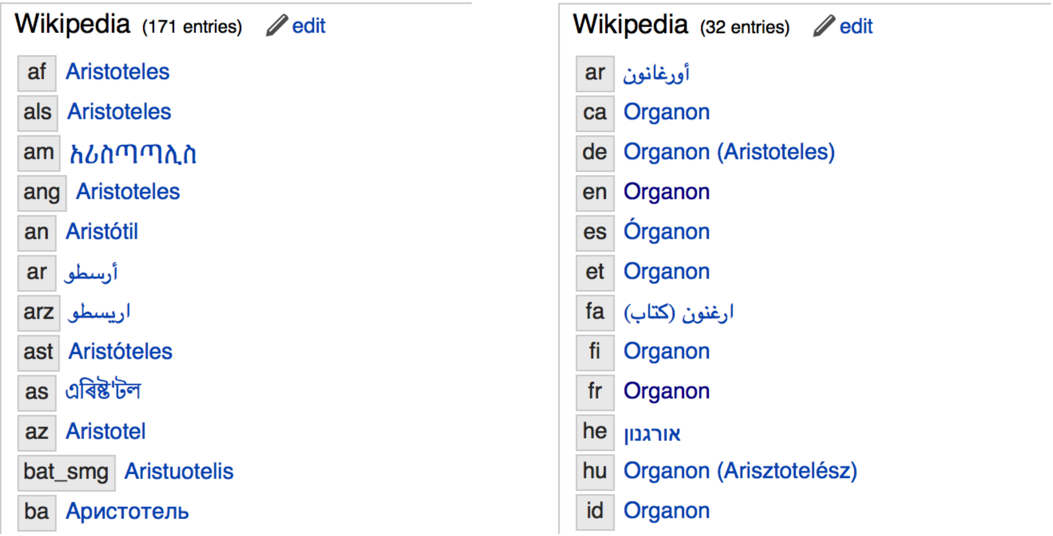


Figure 3: Crosslingual projection of predicate–argument structures

[2] Retrieved on April 7, 2016

Exner et al. (2015) used the Swedish and English versions of Wikipedia to collect a large set of parallel propositions. They identified named entities in these two versions using a unique identifier, the Wikidata Q-number, that enabled them to pair the predicate–argument structures across the languages.

In Wikipedia, a same entity or concept can have from one to more than 200 different language versions. Wikidata is a graph database that connects all these entities and concepts across the versions. It has the form of a centralized repository that stores links to all the versions with a unique number: Q868 in the case of Aristotle and Q500930 for the Organon (Fig. 4).



Wikipedia (171 entries) edit

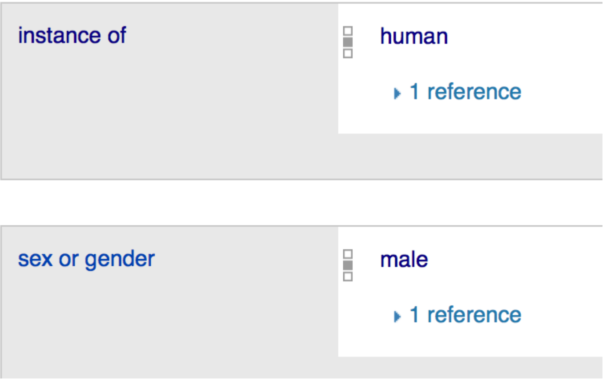
af	Aristoteles
als	Aristoteles
am	አሪስጣጣሊስ
ang	Aristoteles
an	Aristótil
ar	أرسطو
arz	اريسطو
ast	Aristóteles
as	এৰিষ্ট'টল
az	Aristotel
bat_smg	Aristuotelis
ba	Аристотель

Wikipedia (32 entries) edit

ar	أورغانون
ca	Organon
de	Organon (Aristoteles)
en	Organon
es	Órganon
et	Organon
fa	ارغنون (کتاب)
fi	Organon
fr	Organon
he	אורגנון
hu	Organon (Arisztotelész)
id	Organon

Figure 4: Left: The first language versions of Aristotle in Wikidata. The languages in the figure appear in alphabetic order out of 171. Right: The first language versions of the Organon out of 32 entries

In addition to listing entities, Wikidata uses a set of about 2,000 properties[3] to describe them. Aristotle, for example, is an instance of a human (Fig. 5), where *instance of* property, P31, enables the editors to define an ontology.



instance of	human 1 reference
sex or gender	male 1 reference

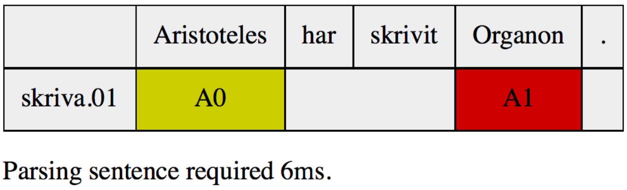
Figure 5: Membership of Aristotle to ontology classes using the *instance of* property

Using the resulting propositions in Swedish, Exner et al. (2015) could train a semantic role labeler. While not on a par with semantic role labelers trained on large hand-annotated corpora, it obtains promising results and in fact can identify the arguments of the Swedish sentence:

Aristoteles har skrivit Organon,
'Aristotle wrote the Organon'

although no such a sentence exists in the Web (Fig. 6).

[3] 1,863 properties as of October 12, 2015



	Aristoteles	har	skrivit	Organon	.
skriva.01	A0			A1	

Parsing sentence required 6ms.

Figure 6: Semantic parsing of *Aristoteles har skrivit Organon*. The arguments are given using the Propbank nomenclature (Palmer et al., 2005), where A0 is the *writer* and A1 is the *thing written*

4 Scaling to Large Corpora

While small corpora can be handled in the form of files, the size of Wikipedia requires a different data structure. We created a document model to store large collections of text. We designed it so that we could store annotations such as token, sentence, paragraph, etc., keep the wiki markup, as well as the subsequent linguistic annotations. Figure 7 shows the structure of this model, Docforia, that consists of multiple layers as well as the conversion process from Wikimedia dumps (Klang and Nugues, 2016). The ideas behind Docforia are similar to those of the UIMA project (Ferrucci and Lally, 2004), but the focus is on simplicity and ease of integration.

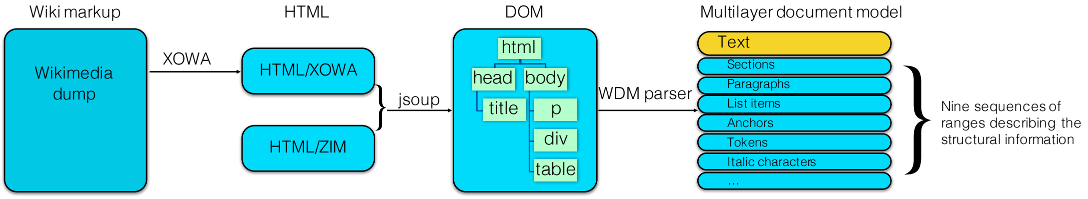


Figure 7: Docforia: A multilayer document model. After Klang and Nugues (2016)

The source code of Docforia as well as its documentation are available from `https://github.com/marcusklang/docforia`.

5 Extraction of Career Timelines

Finally, the retrieval of career timeline is an example of application of knowledge extraction from Wikipedia. Using the Swedish version and the Wikidata ontology, Dib et al. (2015) could extract the careers of people. They restricted the pages to people and analyzed the first paragraph with a dependency parser to find grammatical links between the person described by the page and professions defined as subclasses the Wikidata occupation property. Figure 8 shows an example of sentence parsing to link Göran Persson, a Swedish Prime minister, to *politiker* 'politician' and *statsminister* 'Prime Minister'.

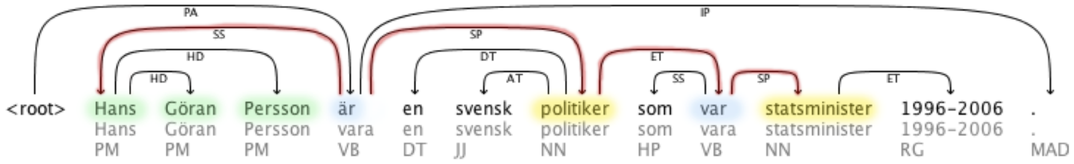
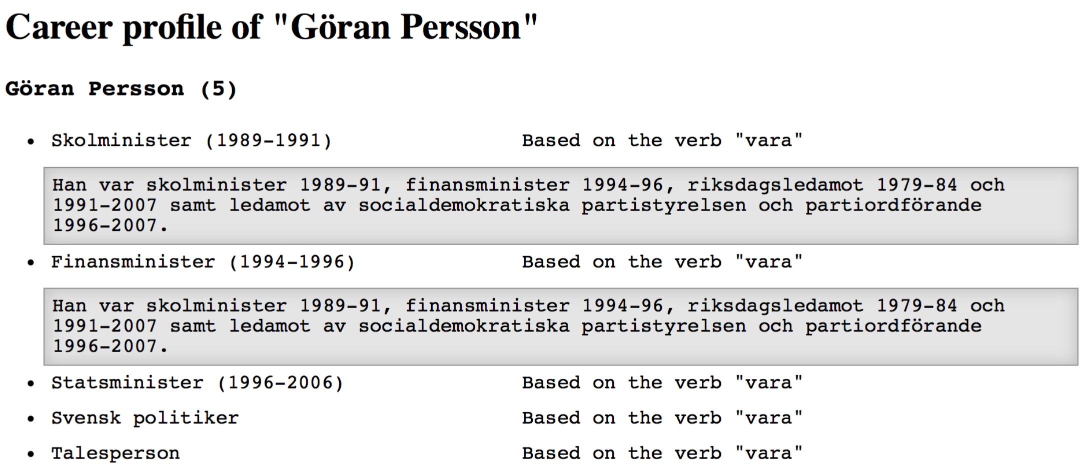


Figure 8: Linking a person to occupations through a dependency graph. After Dib et al. (2015)

Figure 9 shows a screenshot of the application interface.



Career profile of "Göran Persson"

Göran Persson (5)

- Skolminister (1989-1991)   Based on the verb "vara"

```
Han var skolminister 1989-91, finansminister 1994-96, riksdagsledamot 1979-84 och
1991-2007 samt ledamot av socialdemokratiska partistyrelsen och partiordförande
1996-2007.
```

- Finansminister (1994-1996)   Based on the verb "vara"

```
Han var skolminister 1989-91, finansminister 1994-96, riksdagsledamot 1979-84 och
1991-2007 samt ledamot av socialdemokratiska partistyrelsen och partiordförande
1996-2007.
```

- Statsminister (1996-2006)   Based on the verb "vara"
- Svensk politiker   Based on the verb "vara"
- Talesperson   Based on the verb "vara"

Figure 9: The career timeline of Göran Persson

Acknowledgements

This research was supported by Vetenskapsrådet, the Swedish research council, under the *Det digitaliserade samhället* program.

References

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstration Volume*, pages 33–36, Beijing, August 23-27. Coling 2010 Organizing Committee.

Firas Dib, Simon Lindberg, and Pierre Nugues. 2015. Extraction of career profiles from Wikipedia. In *BD2015, Proceedings of the First Conference on Biographical Data in a Digital World 2015*, pages 33–38, Amsterdam, April. CEUR Workshop Proceedings.

Peter Exner and Pierre Nugues. 2012. Constructing large proposition databases. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC 2012)*, pages 3836–3840, Istanbul, May 23–25.

Peter Exner and Pierre Nugues. 2014. REFRACTIVE: An open source tool to extract knowledge from syntactic and semantic relations. In *Proceedings of LREC 2014, The 9th edition of the Language Resources and Evaluation Conference*, pages 2584–2589, Reykjavik, May 27-29.

Peter Exner, Marcus Klang, and Pierre Nugues. 2015. A distant supervision approach to semantic role labeling. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 239–248, Denver, Colorado, June. Association for Computational Linguistics.

David Ferrucci and Adam Lally. 2004. Uima: An architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10(3-4):327–348, September.

David Angelo Ferrucci. 2012. Introduction to “This is Watson”. *IBM Journal of Research and Development*, 56(3.4):1:1 –1:15, May-June.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280:20–32.

Jakob Grundström and Pierre Nugues. 2014. Using syntactic features in answer reranking. In *Proceedings of the AAAI 2014 Workshop on Cognitive Computing for Augmented Human Intelligence*, pages 13–19, Québec, July 27.

Richard Johansson and Pierre Nugues. 2008. Dependency-based semantic role labeling of PropBank. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 69–78, Honolulu, October 25-27.

Marcus Klang and Pierre Nugues. 2016. Wikiparq: A tabulated Wikipedia resource using the Parquet format. In *Proceedings of 10th edition of the Language Resources and Evaluation Conference*, Portorož, May.

2013. Kvitt eller dubbelt – tiotusenkronorsfrågan. `http://en.wikipedia.org/wiki/Kvitt_eller_dubbelt`.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: an annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

Juri Pyykkö, Rebecka Weegar, and Pierre Nugues. 2014. Passage retrieval in a question answering system. In *Proceedings of the The Fifth Swedish Language Technology Conference (SLTC 2014)*, Uppsala, November 13-14.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, and Christopher R. Johnson. 2005. Framenet: Theory and practice. `http://framenet.icsi.berkeley.edu/book/book.html`. Cited 28 October 2005.

Karin Thorsvad and Hasse Thorsvad. 2005. Kvitt eller dubbelt.

Building a Sentiment Lexicon for Swedish

Bianka Nusko
Dept of Philosophy, Linguistics
and Theory of Science,
University of Gothenburg
gusnusbi@student.gu.se

Nina Tahmasebi
Språkbanken,
University of Gothenburg
nina.tahmasebi@gu.se

Olof Mogren
Dept of Computer Science
and Engineering,
Chalmers University of
Technology
mogren@chalmers.se

Abstract

In this paper we will present our ongoing project to build and evaluate a sentiment lexicon for Swedish. Our main resource is SALDO, a lexical resource of modern Swedish developed at Språkbanken, University of Gothenburg. Using a semi-supervised approach, we expand a manually chosen set of six core words using parent-child relations based on the semantic network structure of SALDO. At its current stage the lexicon consists of 175 seeds, 633 children, and 1319 grandchildren.

1 Introduction

Sentiment lexicons have proven a valuable resource for opinion mining tasks. With large amounts of data readily available on the Internet, gathering user sentiments and opinions is a relatively effortless and inexpensive undertaking. As an example, it is possible to easily check whether or not a product is positively received. This information is useful to both potential new customers and the manufacturers or the suppliers of said product. While customers may wish to inform themselves whether the product lives up to the desired quality or value, the company has the opportunity to quickly gather the general consensus on the product and is therefore able to react accordingly. Other examples include politicians or political parties that are able to quickly gather their voters' opinions by using opinion mining.

So far, this field of research is mainly restricted to anglophone data and hence the majority of sentiment lexicons in existence today are in English. Some of the most well-known lexicons include SentiWordNet (Esuli and Sebastiani, 2006), the Bing Liu Opinion Lexicon (Hu and Liu, 2004) and the General Inquirer (Stone et al., 1966). In recent years there has been increasing interest in building opinion lexicons for other languages as well. For German, Remus et al. (2010) built the SentimentWortschatz Lexicon, short SentiWS, using semi-automatic translations of English sentiment resources combined with information about word co-occurrences and word collocations. Banea et al. (2008) use raw data and a bootstrapping method to construct a subjectivity lexicon for languages with scarce resources such as Romanian and Wan (2009) exploits the large amount of annotated English data available to classify Chinese reviews. The approach is supervised and allows classification without any annotated Chinese data.

To the best of our knowledge there are no publicly available lexical resources for sentiment analysis for the Swedish language. Our goal is therefore to lay the groundwork for a Swedish sentiment lexicon.

2 Related Work

Strategies for creating a sentiment lexicon range from purely manual, over semi-supervised, to more or less automatic machine learning approaches. Well known manually constructed sentiment lexicons are e.g. MPQA (Wiebe et al., 2005) and Bing Liu's Opinon Lexicon (Hu and Liu, 2004). However, manual approaches are expensive and difficult to adapt to new domains, therefore, we refrain from using a manual approach and instead rely on a manually created resource.

Severyn and Moschitti (2015) use hashtags and emoticons contained in a Twitter corpus as sentiment information to automatically construct an opinion lexicon. An SVM classifier was trained using words as well as multi-word expressions taken from a distantly-supervised corpus of tweets as features, a corpus that was obtained by using hashtags or emoticons as sentiment indicators. The authors state that the corpus size compensates for the noisiness of the data. While previous approaches like Mohammad et al. (2013) used statistical measures such as PMI to calculate word sentiment association for building the lexicon, Severyn and Moschitti (2015) rely on machine learning techniques and as a consequence achieve higher results. The drawback is that the created lexicons are not necessarily human-interpretable, they include good features for classification and are not meant as stand-alone sentiment lexicons.

A semi-supervised approach that links conjunctions to infer opinion was introduced by Hatzivassiloglou et al. (1997). They used a labelled seed set of adjectives taken from the 1987 Wall Street Journal (WSJ) corpus provided by the ACL Data Collection Initiative (`https://catalog.ldc.upenn.edu/LDC93T1`) and expanded them to pairs of adjectives that are linked by the conjunctions *and* or *but*. Adjectives linked by *and* were assigned the same polarity and adjectives linked by *but* were assigned opposing polarity. A supervised log-linear-regression model was used for this task.

Esuli and Sebastiani (2005), who were involved in building SentiWordNet, used textual term descriptions called *glosses* to determine the polarity of a sense by assuming that, generally, terms with similar glosses also have similar polarity. The method is used to expand a set of seed words by means of the lexical relations specified in a thesaurus. For each term in the expanded set the gloss is extracted from an online dictionary. After conversion into a vectorial format, a binary classifier is trained on this data and finally applied to a new data set. An advantage of this method is that it is not restricted to classifying adjectives, like Hatzivassiloglou et al. (1997)'s approach, but allows classification of all terms. A disadvantage is that the method can produce noisy results.

3 The Resource: SALDO

SALDO (Borin et al., 2013) is a publicly available electronic resource for the written modern Swedish language. It was developed and is continuously expanded at Språkbanken, the Swedish Language Bank, at the University of Gothenburg. SALDO includes semantic as well as morphological information about words. It is organised as a lexical-semantic network linking word senses by their association to other word senses and does not purely rely on synonymy relations like the synset structure of Princeton WordNet (`http://wordnet.princeton.edu`) for example. Furthermore, SALDO contains words from all word classes, including closed-class words, although the full scope of the word classes are not exploited for the sentiment lexicon described in this paper.

All entries, i.e. word senses, in SALDO are arranged hierarchically around the dummy core *PRIM*. Each word sense, except the dummy, has a primary semantic descriptor and optionally one or more secondary semantic descriptors – all of which are word senses in SALDO as well. While some word senses listed in SALDO have more than one secondary descriptor, most of them being names, this does not apply to the words included in our lexicon. Therefore, there is no strategy for dealing with words that have two or more secondary descriptors. In this study we use the semantic descriptors assigned to a word sense to determine the sentiment of that sense. Additionally, we also extract the lemgram and part-of-speech information for each word sense from SALDO.

primary descriptor / secondary descriptor The primary descriptor "(1) [...] is a semantic neighbor of the entry to be described and (2) it is more central than it" (Borin et al., 2013, p. 1195). Secondary descriptors give more information about the particular sense of the word and can also include words that negate or indicate the strength of the primary descriptor. Therefore they can also modify the polarity of a word[1]. The descriptors link all the words in SALDO to form a network.

[1] Due to the modifying qualities of certain words, we specifically removed *lagom* "just the right amount" and all words that have *sjukdom* "disease" in them from our data sets to reduce noise.

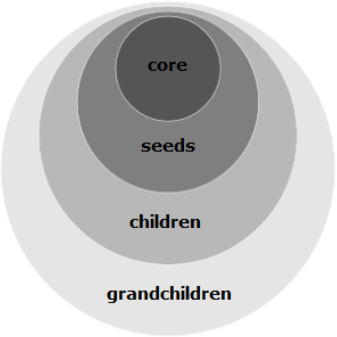


Figure 1: structure of the data sets.

lemgram "A *Lemgram* in SALDO is a combination of a base form and an inflectional pattern" (Borin et al., 2013, emphasis added). That is, a lemgram can basically be seen as denoting an inflection table. It should not be confused with *lemma*. Although lemma can refer to the headword of a set of inflectional forms, which is very similar to the notion of lemgram, in most cases it simply connotes the base form of a word.

4 Methodology

Similarly to Esuli and Sebastiani (2005), we start with a small set of core words that is expanded semi-automatically using the lexical-semantic relations of the network structure in SALDO. Figure 1 shows how our data sets are structured. We first introduce the format of our data in section 4.1 and explain the expansion process in the subsequent section 4.2.

4.1 Format

Each entry in the data set is represented in the format *(word, polarity, strength, word sense, written form(s), part-of-speech, confidence, lemgram frequency, example)*.

polarity We manually assign a polarity to the six core words. The polarity of the seed words then depends on the polarity of the core word that is its primary descriptor. The same is true for the children and grandchildren which are assigned the polarity of their primary descriptor as well. In cases where the secondary descriptor is *inte* "not" or *motsats* "the opposite" the polarity of a word is reversed.

strength The polarity strength is a discrete number in the range [1,4] and [-4,-1]. Positive numbers indicate positive word polarity and negative numbers express negative polarity. All words are initially assigned a value of 2 or -2 respectively. A higher number indicates a stronger polarity. Certain secondary descriptors can increase or decrease these values so that words with one of the secondaries shown in Table 1 get the respective strength as specified below.

lite "a little"	1, -1
none	2, -2
mycket "very"	3, -3
enastående "outstanding", värdelös "worthless"	4, -4

Table 1: secondary descriptors and corresponding strength values.

word form(s) and word sense A word sense is the semantic sense or meaning of a word recorded in SALDO. A sense is written in the form *wordsense..sensenumber*, i.e. *bra..1*. The word form is the written form of the word sense, i.e *bra*.

part-of-speech (pos) The part-of-speech tag of a word extracted from SALDO. We only use words that are either adjectives, adverbs or verbs, i.e having one of the following tags: *av, avh, avm, ava, vb, vbn, vba, ab*. All other words, especially nouns, were removed since they did not add polarised words to our data.

confidence The confidence score, ranging [0,1], indicates how certain we are that the assigned polarity is correct. The higher the score, the more confident we are in the polarity. Seed words get a value of 1, children a value of 0.75 and grandchildren a value of 0.5. So the value decreases the further away we get from the seed words.

lemgram frequency Using a large data set of modern Swedish, the Gigaword corpus (Eide et al., 2016), we measure the frequency of the lemgram corresponding to the word sense.

example For each word we intend to provide an example sentence taken from the same data set used for computing the above frequencies. This will allow us to examine the context in which the word occurs. Deciding on how to extract the most useful example sentence for each entry in our lexicon is future work and will be our next focus.

4.2 Expanding the Seed Set into a Sentiment Lexicon

The core of our data set consists of the six words *bra* "good", *glad* "happy", *frisk* "healthy", *dålig* "bad", *ledsen* "sad" and *sjuk* "ill". We expand this core set into the different seed sets shown in Figure 1 as described below:

1. Extract all words from SALDO that have one of the core words as their primary descriptor.
2. Extract and add children, i.e. words having one of the seed words as primary descriptor.
3. Extract and add grandchildren, i.e. words having one of the children as their primary descriptor.

A minimal expansion example is the following (The two underlined items are the primary and secondary descriptor of the word in first position of the entry.):
core word: *bra*
seed: *(god, +, 2, god..1, ['god'], bra..1, None, av, 1, LEMGRAMFREQ, EXAMPLE)*
child: *(elak, -, -2, elak..1, ['elak'], god..1, inte..1, av, 0.75, LEMGRAMFREQ, EXAMPLE)*
grandchild: *(sarkastisk, -, -2, sarkastisk..1, ['sarkastisk'], elak..1, None, av, 0.5, LEMGRAMFREQ, EXAMPLE)*

The polarity, its strength, and also the certainty value of a word sense depend on the values of its primary and secondary descriptors. We add *god* "kind" to our set of seed words because it has the core word *bra* "good" as a primary descriptor. A child of *god* is *elak* "mean" with *god* itself as its primary and the negation *inte* "not" as its secondary descriptor. Because of this secondary descriptor we give the word a negated polarity. *Elak* in turn has *sarkastisk* "sarcastic" as a child. The polarity of this child is the same as the polarity of its parent because it lacks a secondary descriptor that negates its polarity. Being the child of *elak*, *sarkastisk* is at the same time the grandchild of *god*.

5 The Sentiment Lexicon and Its Quality

The seeds extracted for the core words amounted to 233 words in the seed set. The two expansion steps yielded 1344 children and finally 3848 grandchildren. After removing all words that have part-of-speech tags other than those listed under the *part of speech* section, the sentiment lexicon consists of 175 seeds, 633 children, and 1319 grandchildren.

To test the quality of the sentiment lexicon, we performed an evaluation on 100 words, 50 positive and 50 negative, from each of our three seed sets. All 300 words were shuffled and manually labelled by at least two native speakers of Swedish. Since determining the overall sentiment of a word is easier than deciding on its strength, we only evaluated the polarity. The annotators were asked to assign each word to one of the following four categories: positive, negative, neutral or unknown. A Fleiss' Kappa (Fleiss, 1971) score of 0.70[2] indicates *substantial agreement* (Artstein and Poesio, 2008) between the annotators.

Out of the 150 words that were evaluated by all three evaluators, 113 were fully agreed upon. For 107 of them, the manually assigned polarity corresponds to the respective sentiment listed in our lexicon giving us a precision of 71%. Partial agreement, i.e. at least two evaluators assigned the same label to a word, was reached in 145 cases. 128 out of these are in accordance with our automatically assigned labels, resulting in a precision of 85%. While one of the annotators classified 13 words as having unknown polarity, each of the remaining two evaluators only marked 5 words words as unknown. Although the evaluation showed that parts of the data were difficult to classify, the agreement score $\kappa = 0.70$ and the low number of words labelled as unknown show that the annotators had a good intuition about word sentiment in the majority of cases. The polarity of the 22 incorrect words (as agreed by at least two annotators) has been reversed in the sentiment lexicon.

6 Outlook

Currently, the entries of our lexicon are not supported by example sentences. An important task is therefore to find a representative example for each entry which shows the usage as well as the context of the respective word sense and can be used for e.g. classification. These sentences should also serve as a source for extracting possible new words to further expand our data set beyond SALDO. Since the final lexicon should be genre-independent, we must ensure that the extracted examples are representative for the general language or include examples for different genres such as newspapers and social media.

A future step is to refine the strategies for assigning confidence scores and polarity strength. Confidence is assigned using a heuristics based on the word's placement in the parent-child-grandchild hierarchy but at present does not extend beyond words included in SALDO. Determining basic polarity is, for the time being, a binary decision between the discrete values *positive* (plus sign) and *negative* (minus sign). A separate strength value is given to supplement the polarity and further differentiate between the words within the same category. The main reason for this is that strength is much harder to determine than the basic polarity and current strength assignment can be seen as an approximation at best. Also, since we have more confidence in the specified polarities than the respective strength, keeping these values separated simplifies working with the data. In the current version of the lexicon, only words listed in SALDO will get a strength value and it is future work to find a method for also assigning strength to words that cannot be found in the resource.

The data set is available under the following link:
`https://spraakbanken.gu.se/swe/resurs/sentimentlex/xml`

Acknowledgements

We would like to thank Richard Johansson and Markus Forsberg for sharing ideas and expertise. The authors would also like to acknowledge the project “Towards a knowledge-based culturomics” supported by a framework grant from the Swedish Research Council (2012– 2016; dnr 2012-5738).

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Carmen Banea, Janyce M Wiebe, and Rada Mihalcea. 2008. A bootstrapping method for building subjectivity lexicons for languages with scarce resources.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. Saldo: a touch of yin to wordnets yang. *Language resources and evaluation*, 47(4):1191–1211.

[2] Since one of the annotators only labelled the first 150 items in the evaluation data, the agreement calculation is based on this set of words.

Stian Rødven Eide, Nina Tahmasebi, and Lars Borin. 2016. The Swedish Culturomics Gigaword Corpus: A One BillionWord Swedish Reference Dataset for NLP. In *From Digitization to Knowledge 2016, D2K16 in conjunction with Digital Humanities 2016.*

Andrea Esuli and Fabrizio Sebastiani. 2005. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 617–624. ACM.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Vasileios Hatzivassiloglou and Kathleen R McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, pages 174–181. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242.*

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. Sentiws-a publicly available german-language resource for sentiment analysis. In *LREC.*

Aliaksei Severyn and Alessandro Moschitti. 2015. On the automatic learning of sentiment lexicons. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2015).*

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 235–243. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Formalizing and Querying a Diachronic Termino-Ontological Resource: the *CLAVIUS* Case Study

Silvia Piccini, Andrea Bellandi, Giulia Benotto
Istituto di Linguistica Computazionale “Antonio Zampolli”,
Via Giuseppe Moruzzi 1, 56124 Pisa, Italy
name.surname@ilc.cnr.it

Abstract

In this work, we describe the modelling of a diachronic termino-ontological resource, named *CLAVIUS*, representing the evolution of astronomical concepts and theories from antiquity until the dawn of the modern age. The resource was built by means of existing tools allowing the scholars to formalize knowledge even though they are not familiar with the models and the languages underlying the representation. More specifically, *Protégé*, a free open-source ontology editor, which supports OWL (and OWL 2) and *Chronos*, a plug-in for *Protégé* to manage temporal aspect, were used. A raw evaluation of the resource is provided by means of a controlled natural language interface, which enables scholars to answer a set of salient queries defined by our domain expert.

1 Introduction

Concepts, as well as the words used to evoke them, are subject to the inexorable law of change. This is particularly the case for the history of science. Over the centuries scholars have built different theoretical models in response to the continuous innovation that emerged from observation sometimes producing a real scientific revolution in the world view. Needless to say, a change in conceptual level often corresponds to a change in terminology: new terms can be introduced to express the new system of concepts while old terms can be dismissed when the concept becomes obsolete or can refer to new concepts.

Technology can come to the aid of scholars in their endeavors with textual hermeneutics and information retrieval. The key point is that the concepts evoked within the text, as well as the terms representing these concepts, need to have a structured organization, and to be explicitly and univocally defined through the relationships that unite them.[1] In other words, the knowledge conveyed by a text needs to be represented in a termino–ontological resource where the ontology is connected to a lexical component.

In the age of the Semantic Web it is important that these resources are built according to the technologies of Semantic Web and Linked Open Data, in order to enable interoperability so that they can be shared, and reused across scientific communities (Ciotti, 2014).

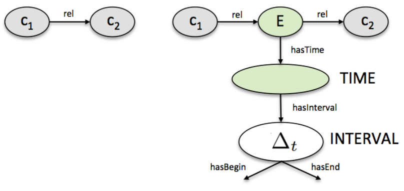
In this work we present the modelling of a diachronic termino-ontological resource, named *CLAVIUS*, devoted to the astronomy domain in a time span ranging from antiquity until the dawn of the modern age. *CLAVIUS* was built within the Project Clavius on the Web[2] and got its name from Cristophorus Clavius, a German Jesuit mathematician and astronomer, who was one of the most respected and influential scholars of his time. The texts he wrote were widely used within Europe to teach astronomy and were studied even by scholars such as René Descartes, Marin Mersenne and Johannes Kepler (Lattis, 1994).

The resource was built on the basis of Clavius’ commentary on Sacrobosco’s *De sphaera mundi*, both of which are considered highly influential works of pre-Copernican astronomy in Europe. In his 500-hundred page tome the Jesuit father describes and comments on the tenets of ancient and medieval astronomy, most often upholding the traditional standpoint. In the commentary, the history of scientific thought is described from Platonic-Aristotelian theories to the early *novitates astronomicae*, that pave

[1] Although the conceptual and the terminological layers are intimately linked, the theoretical necessity of distinguishing between them led to the development of new paradigms (Roche, 2007), and strategies (Reymonet et al., 2007).

[2] *http://claviusontheweb.it/*

Relation	Temporalization
$c_1\ rel\ c_2$	$\exists\Delta_t.\ c_1\ rel\ E, E\ rel\ c_2$



Property	Inferences
Reflexive	$\exists\Delta_t.\ E\ rel\ E, c_1\ rel\ c_1, c_2\ rel\ c_2$
Symmetric	$\exists\Delta_t.\ E\ rel\ c_1, c_2\ rel\ E$
Transitive	$\forall\Delta_t.\ c_1\ rel\ c_2$
Inverse	$\exists\Delta_t.\ E\ rel^{-1}\ c_1, c_2\ rel^{-1}\ E$

Figure 1: Examples of inferences about a relation rel by means of n-ary model.

the way for the modern era and signal the end of the medieval world view.
The remainder of this paper is organised as follows: in Section 2, we will present the main approaches for representing temporal aspects in ontologies; in Section 3 we will describe the *CLAVIUS* resource, and in Section 4, we will present a controlled natural language interface able to answer a set of salient queries defined by our domain expert. Finally, in Section 5 some conclusions will be drawn.

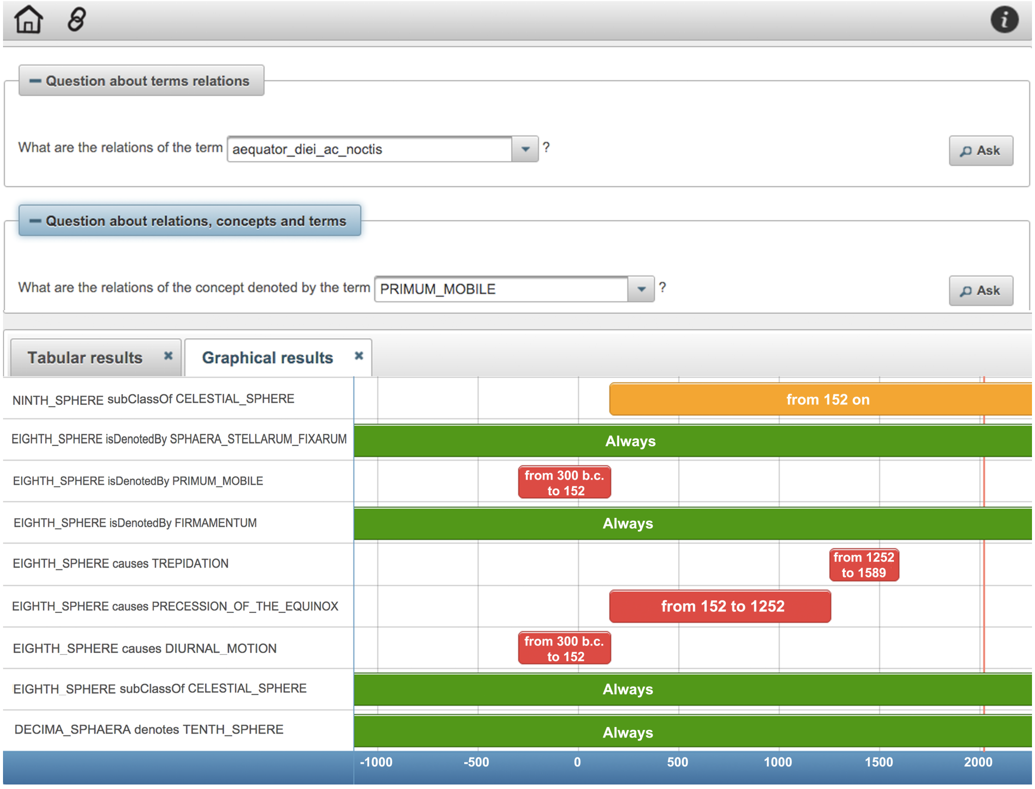
2 Approaches to Represent Temporal Aspects

In his commentary on Sacrobosco's *De Sphaera*, Clavius illustrates the evolution of the authors' conceptualization from the antiquity until the dawn of the modern age. In order for scholars to access the semantic content of the text through "sophisticated" queries, the evolution of astronomical concepts - as well as of the terms adopted to evoke them - need to be formalized into a dynamic and temporal ontology. In literature, the problem of representing the dynamically evolving information in ontologies has been addressed by adopting several different approaches (Flouris et al., 2008). One very simple solution is to create a version of the ontology for each temporal event as described in (Grandi and Scalas, 2009) (ontology versioning). For this, an ad-hoc versioning algorithm is developed in order to access the different temporal variants. Other solutions are the reification approach, the n-ary model, and the 4D-fluent approach. The first suffers from data redundancy (Batsakis and Petra, 2011), and offers limited OWL reasoning capabilities (Welty et al., 2006). The n-ary model represents a relation as two properties, each one related with a new object. These two objects are linked to each other with an n-ary relation. This approach requires only one additional object for every temporal interval, maintains property semantics but suffers from data redundancy in the case of inverse and symmetric properties (Noy and Rector, 2006). Concerning the 4D-fluent approach, described in (Welty et al., 2006), concepts in time are represented as 4-dimensional objects with the 4^{th} dimension being the time. However, all the proposed approaches lead to a massive proliferation of objects, making reasoning and querying unnecessarily complex, expensive, and error-prone. This is due to the underlying data structure, the RDF triple (Krieger, 2010). Consequently, each model presents advantages and disadvantages. The choice of a model is linked to the specific needs for representation, querying and reasoning. In the following, we will show the model adopted in *CLAVIUS* and we will give reasons for our choice.

3 Formalizing the Astronomical Domain in *CLAVIUS*

The two key points determined our approach to modelling the static and dynamic knowledge conveyed in Clavius' texts have been i) the use of standard Semantic Web and Linked Data technologies, and ii) the possibility to use existing tools, which allow the scholars to formalise knowledge even though they are not familiar with the models and the languages underlying the representation.
Concerning i), *CLAVIUS* was coded in OWL (Web Ontology Language), which is the family of knowledge representation languages for authoring ontologies. As regards ii), the resource was built in Protégé, the most well-known free and open source editor of ontologies supporting OWL natively. The diachronic aspect was modelled using $Chronos$ (Preventis et al., 2012), a plug-in for Protégé aimed at managing temporal ontologies. It can be downloaded from the Web, and is based on n-ary relation model.

In *CLAVIUS* ontology concepts and terms were represented both as OWL classes in order to ensure



Question about terms relations

What are the relations of the term aequator_diei_ac_noctis ?

Question about relations, concepts and terms

What are the relations of the concept denoted by the term PRIMUM_MOBILE ?

NINTH_SPHERE subClassOf CELESTIAL_SPHERE	from 152 on
EIGHTH_SPHERE isDenotedBy SPHAERA_STELLARUM_FIXARUM	Always
EIGHTH_SPHERE isDenotedBy PRIMUM_MOBILE	from 300 b.c. to 152
EIGHTH_SPHERE isDenotedBy FIRMAMENTUM	Always
EIGHTH_SPHERE causes TREPIDATION	from 1252 to 1589
EIGHTH_SPHERE causes PRECESSION_OF_THE_EQUINOX	from 152 to 1252
EIGHTH_SPHERE causes DIURNAL_MOTION	from 300 b.c. to 152
EIGHTH_SPHERE subClassOf CELESTIAL_SPHERE	Always
DECIMA_SPHAERA denotes TENTH_SPHERE	Always

Figure 2: An example of answer to the query "*What are the relations of the concept denoted by the term Primum Mobile*" is presented. A demo version is available at: *http://146.48.93.19:8080/clavius*

autonomy of the terminological and the conceptual level. The top level of the ontology is represented by the two OWL disjoint classes CONCEPT and TERM, which subsume, respectively, all astronomical concepts and all astronomical terms. In *CLAVIUS* the conceptual level is expressed in English while the terminological level is made up of Latin words which are linked to the concepts they evoke through the relation *denotes* (and its inverse *isDenotedBy*). The ontology currently consists of 106 classes organised into four hierarchical levels, a set of 10 DataType Properties and 18 Object Properties, which make it possible to give a precise representation of the concepts and the terms. There are three basic types of relations: **lexical**, which express the paradigmatic relations among terms (hypernymy, hyponymy, meronymy, holonymy, synonymy and antonymy); **inter-level** which link the sense of a term to the concept it evokes; **conceptual** which describe the relations holding between the concepts. Among these there are **domain specific** relations, introduced to better formalize the characteristics of the astronomical domain (*isNear*, *revolvesAround*, etc.). Finally, when opportune, the relations were easily temporalized by means of the $Chronos$ editor. However, as described in Section 2, adopting the n-ary model brings about issues of reasoning. In fact, uncorrected or redundant triples are inferred, w.r.t. the temporalized properties, in particular when these properties are defined as symmetric, transitive or they have an inverse. In Figure 1 an example of inferences referred to a relation rel is shown. To solve such problems, it is necessary to filter out from the SPARQL queries results all the triples that were erroneously inferred or redundant. For example, as Figure 1 depicts, if rel holds in a particular time frame and is transitive, a reasoner infers that rel always holds in time.

4 Querying *CLAVIUS*

In order to give a raw evaluation of the formalised resource as well as to facilitate access for scholars, a controlled natural language interface (Schwitter, 2010) was developed to query the ontology. As illustrated in Figure 2, query templates were created, and each of them is made up of a fixed part that typifies a specific querying model and a variable part that allows the user to select an element of the ontology from

the drop-down list. Question templates are processed by the software into SPARQL queries. Queries can be made in controlled natural language by taking into account the lexical level (*"What are the relations of a specific term?"*), the ontological level (*"What are the relations of a concept denoted by a specific term?"*) or both. The questions themselves could involve diachronic aspects, such as in *"What relation exists between two concepts in a specific temporal interval?"*.
In Figure 2 an example of query is provided. The term *primum mobile* is highly ambiguous as over time it has denoted different concepts. In the Aristotelian view it was the eighth sphere, the outermost sphere containing the fixed stars, also called *sphaera stellarum fixarum* or *firmamentum*. With Ptolemy, *primum mobile* became the ninth sphere, introduced to explain the precession of the equinoxes observed by Hipparchus. In the Alfonsine Tables *primum mobile* denoted the tenth sphere, which was added to explain the trepidation motion noted by the astronomer Thabit Ibn Qurra Arab [3]. In Figure 2 it is possible to see how the relations between the term *primum mobile* and the concepts it denotes have changed over time.

5 Conclusions

Formalizing the evolving knowledge conveyed by Clavius' work raised interesting challenges about knowledge representation. From a theoretical standpoint, many models have been proposed to formalize the diachronic evolution of information. Nevertheless, either these models are not supported by scholar-oriented tools or the available tools are based on approaches which lead to a massive proliferation of objects, making reasoning and querying complex and error-prone. This paper is intended as a springboard to discussion within the scientific community, in particular Digital Humanities, which increasingly feels the need to adopt Semantic Web technologies in order to create resources that can be shared, reused and built on by scholars.

References

Batsakis S., Petrakis, E.G.M. 2011. Representing Temporal Knowledge in the Semantic Web: The Extended 4d Fluents Approach. *Combinations of Intelligent Methods and Applications*, pp. 55–69, Springer, Heidelberg.

Ciotti, F. 2014. Digital Literary and Cultural Studies: State of the Art and Perspectives. *Between*, vol. 4, no. 8.

Flouris, G., Manakanatas D., Kondylakis H., Plexousakis D., Antoniou G. 2008. Ontology Change: Classification and Survey. *The Knowledge Engineering Review*, vol. 23, pp. 117–52, Cambridge University Press.

Grandi, F., Scalas M.R. 2009. The Valid Ontology: A Simple OWL Temporal Versioning Framework. 3^{th} *International Conference on Advances in Semantic Processing*, pp. 98–102, IEEE Press, New York.

Krieger, H.U. 2010. A General Methodology for Equipping Ontologies with Time. *Proceedings of LREC 2010.*

Lattis J. 1994. Between Copernicus and Galileo: Christoph Clavius and the Collapse of Ptolemaic Cosmology. *University of Chicago Press*, Chicago.

Noy N., Rector B. 2006. Defining N-ary Relations on the Semantic Web. *W3C Working Group Note*, http://www.w3.org/TR/swbpn-aryRelations

Preventis A., Marki P., Petrakis E.G.M., Batsakis S. 2012. Chronos: A Tool for Handling Temporal Ontologies in Protégé. 24^{th} *International Conference on Tools with Artificial Intelligence*, Athens, Greece.

Reymonet A., Thomas J., Aussenac-Gilles N. 2007. Modelling ontological and terminological resources in OWL DL. *The Lexicon/Ontology Interface, Workshop*, pp. 415–425, Busan (South Korea).

Roche C. 2007. Le terme et le concept: fondements d'une ontoterminologie. *Actes de la première conférence TOTh*, pp.1–22, Annecy.

Schwitter R. 2010. Controlled Natural Languages for Knowledge Representation. 23^{rd} *International Conference on Computational Linguistics*, pp. 1113–1121.

Welty, C., Fikes R., Makarios S. 2006. A Reusable Ontology for Fluents in OWL. 4^{th} *International Conference FOIS*, pp. 98–102, IEEE Press, New York.

[3] C. Clavius, *In Sphaeram Comm.*: *"Quare cum corpus simplex vnico tantum motu ferri sit aptum, ut uolunt philosophi, non potest nonum coelum esse primum mobile, sed supra ipsum erit decimum statuendum coelum, quod sit primum mobile"*.

ParaViz, an online visualization tool for studying variation in meaning based on parallel texts

Ruprecht von Waldenfels
Dept. of Slavic languages and literatures
University of California, Berkeley
ruprecht.waldenfels@gmail.com

Michał Woźniak
Institute of Polish
Polish Academy of Sciences, Cracow
michal.wozniak@ijp-pan.krakow.pl

Abstract

ParaViz is a modular corpus query and analysis tool for use with a word aligned, linguistically annotated multilingual corpus of parallel translated texts. Representing an addition to classic query-based corpus tools, ParaViz makes it easy to assess differences in the meanings of cognate or otherwise comparable items in different languages based on their distribution in parallel texts. Translations are thus essentially used as semantic annotations, allowing for a bottom-up analysis of semantics in a network of texts in many languages.

The tool takes as input a user-supplied operationalization of the variables under comparison. It then provides the user with two perspectives on the distribution of these variables in the parallel corpus: on the one hand, a close-up perspective of word-aligned corpus examples, color-coded in respect to the user-provided parameters; on the other hand, a bird's view perspective with visualizations that provide overviews of the aggregated differences in use. Data sets with the categorized data is made available for download so it can be further analyzed.

Initially developed as an offline version with a specific research topic in mind, the tool has been adapted as an online tool and will be available for use with the ParaSol corpus (Waldenfels 2011). We feel the publication of such tools in a format that makes it accessible for the research community at large is an important part of addressing the issues of research result replication and sustainability of research efforts in digital humanities in general.

1 Introduction

The article reports on work on ParaViz, a complex query and visualization system for word aligned, linguistically annotated parallel corpora used to investigate cross-linguistic similarity of linguistic variables. ParaViz builds on a simple insight: similarities in the distribution of linguistic items in parallel texts, i.e., multiple translations of the same text in different languages, reflect their functional and semantic similarity across languages in a distributional model of semantics (for such models, see the overview in Sahlgren 2008). By comparing such distributions in a word aligned corpus, notions of comparative semantics can be achieved bottom-up; see Cysouw and Wälchli (2007), Dahl (2014) for related approaches, Waldenfels (2015b) for more background and the description of an earlier version of ParaViz.

ParaViz in its offline version is a functional and powerful research instrument developed in a concrete research project that aims to investigate language convergence and divergence based on a parallel corpus (von Waldenfels, 2014). Perhaps typically for such a project, its production version today resembles a patchwork of different technologies and involves many semi-automated steps. It is designed to be used with a locally available parallel corpus – data that is not

```
<parameter id="NounSuffixes">
  <type id="O" name="OST">
    <criteria><lng>ru</lng>
      <regexp level="lem">ость$</regexp><regexp level="tag">^N.*</regexp>
    </criteria>
    <criteria><lng>sl</lng>
      <regexp level="lem">ost$</regexp><regexp level="tag">^S.*</regexp>
    </criteria>
   <criteria><lng>pl</lng>
     <regexp level="lem">ość$</regexp><regexp level="tag">^subst.*</regexp>
    </criteria>
  </type>
  <type id="S" name="STVO">
    <criteria><lng>ru</lng>
      <regexp level="lem">ство$</regexp><regexp level="tag">^N.*</regexp>
    </criteria>
    <criteria><lng>sl</lng>
      <regexp level="lem">stvo$</regexp><regexp level="tag">^S.*</regexp>
    </criteria>
    <criteria><lng>pl</lng>
      <regexp level="lem">[cs]two$</regexp><regexp level="tag">^subst.*</regexp>
    </criteria>
  </type>
</parameter>
```

Figure 1: A sample parameter file, defining classes of cognate suffixes that are defined for each language slightly differently.

unproblematic to share from both a practical and a legal point of view[1]. These facts make it difficult for other researchers to use the methodology and tools that were developed for the original project and make it virtually impossible to replicate its results, both of which would be highly desirable.

The aim in developing an online version of ParaViz is to reuse the existing system to create a web service which takes care of all technicalities and provides users with an easy way to conduct their own research with this corpus. In the design of the system, we aim to significantly lower the threshold for researchers that want to do comparable research, empowering them to use our methods and, replicate, test and expand on our results.

Section two introduces the functions of the tool in some more detail and in the context of ongoing research. Section three presents a description of design choices and the user interface. Section four concludes with an outlook to further developments.

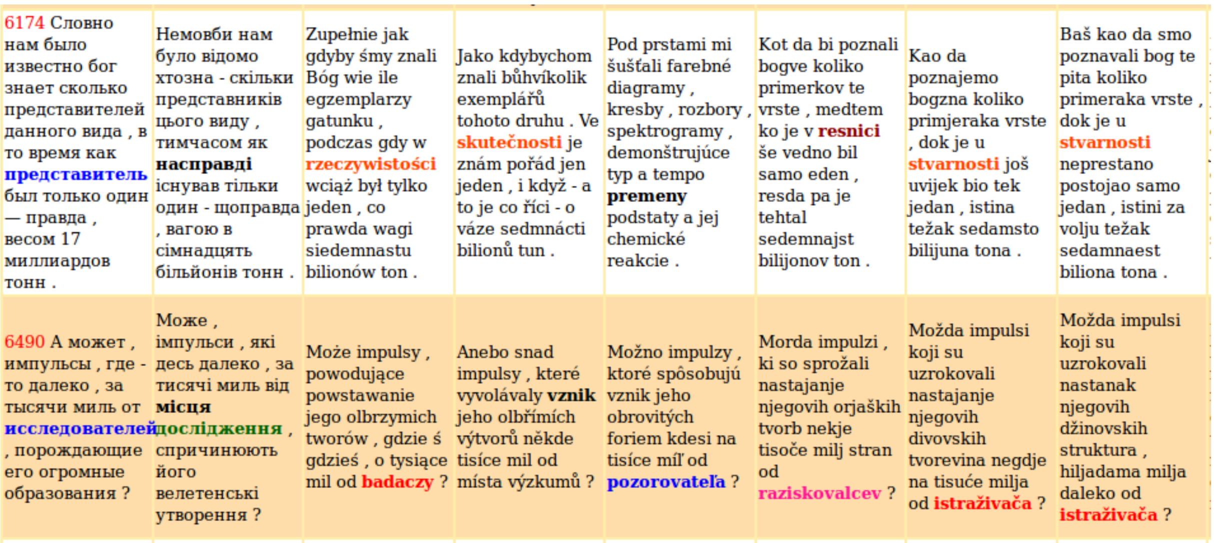
2 Multilingual query and visualization based on parameter files

ParaViz allows the user to define items for comparison using a standardized parameter file in XML format. In this file, the user can define cross-linguistic types the use of which is compared on the basis of word and sentence alignment encoded in the corpus. The types are defined as regular expressions over tokens and their linguistic annotation, which at the moment of writing involves morphosyntactic tagging and lemmatization; in the future, semantic annotation may be added.

As an example, figure 2 shows a set of parameters that defines the cognate suffix classes OST and STVO, both forming abstract nouns, in three Slavic languages. These suffixes are used in all Slavic languages for the derivation of abstract nouns, e.g., Russian *molod-ost'* 'youth', Polish *rad-ość* 'happiness', Serbian *mogućn-ost* 'novelty' all represent instances of the same cognate suffix 'OST'. It is represented in slightly different forms and with slightly different usage profiles across the Slavic languages. In the original project, over ten such cognate suffix classes are defined.

The system then provides the user with two representations of the data that result from

[1] For the relevance of this point in the Swiss legal context, see www.swisscorpora.ch.



6174 Словно нам было известно бог знает сколько представителей данного вида , в то время как **представитель** был только один — правда , весом 17 миллиардов тонн .	Немовби нам було відомо хтозна - скільки представників цього виду , тимчасом як **насправді** існував тільки один - щоправда , вагою в сімнадцять більйонів тонн .	Zupełnie jak gdyby śmy znali Bóg wie ile egzemplarzy gatunku , podczas gdy w **rzeczywistości** wciąż był tylko jeden , co prawda wagi siedemnastu bilionów ton .	Jako kdybychom znali bůhvíkolik exemplářů tohoto druhu . Ve **skutečnosti** je znám pořád jen jeden , i když - a to je co říci - o váze sedmnácti bilionů tun .	Pod prstami mi šušťali farebné diagramy , kresby , rozbory , spektrogramy , demonštrujúce typ a tempo **premeny** podstaty a jej chemické reakcie .	Kot da bi poznali bogve koliko primerkov te vrste , medtem ko je v **resnici** še vedno bil samo eden , resda pa je tehtal sedemnajst bilijonov ton .	Kao da poznajemo bogzna koliko primjeraka vrste , dok je u **stvarnosti** još uvijek bio tek jedan , istina težak sedamsto bilijuna tona .	Baš kao da smo poznavali bog te pita koliko primeraka vrste , dok je u **stvarnosti** neprestano postojao samo jedan , istini za volju težak sedamnaest biliona tona .
6490 А может , импульсы , где - то далеко , за тысячи миль от **исследователей** , порождающие его огромные образования ?	Може , імпульси , які десь далеко , за тисячі миль від **місця** **дослідження** , спричинюють його велетенські утворення ?	Może impulsy , powodujące powstawanie jego olbrzymich tworów , gdzie ś gdzieś , o tysiące mil od **badaczy** ?	Anebo snad impulsy , které vyvolávaly **vznik** jeho olbřímích výtvorů někde tisíce mil od místa výzkumů ?	Možno impulzy , ktoré spôsobujú vznik jeho obrovitých foriem kdesi na tisíce míľ od **pozorovateľa** ?	Morda impulzi , ki so sprožali nastajanje njegovih orjaških tvorb nekje tisoče milj stran od **raziskovalcev** ?	Možda impulsi koji su uzrokovali nastajanje njegovih divovskih tvorevina negdje na tisuće milja od **istraživača** ?	Možda impulsi koji su uzrokovali nastanak njegovih džinovskih struktura , hiljadama milja daleko od **istraživača** ?

Figure 2: A word aligned corpus sample with color coding according to user-supplied parameter file.

applying these parameters as queries in the corpus. First, it produces random samples of relevant corpus examples with the respective aligned word forms given in bold and in different colors according to the criteria in the user supplied parameter file. This allows users qualitative insight into the data and lets them gauge the error rate of the operationalization; see figure 2.

Second, the co-occurrence patterns found in the corpus are visualized as NeighborNets: figure 3 shows two such graphs. The left graph represents similarities and differences in the distribution of OST across 14 Slavic versions of the same text, partly in multiple translations. It shows that the use of the suffix in duplicate translations, e.g., Polish and Polish/2, is very similar, and between different languages, it mostly follows the accepted division of languages into East, West and South Slavic. However, there is an important exception: Russian and Bulgarian cluster together, showing that use of this suffix in these two languages is very similar due to extensive language contact in the history of these languages.

The right graph gives an overview of the similarity of cognate suffix classes across all Slavic languages: here, all the suffixes are compared in relation to how often they are used in equivalent word forms across different Slavic versions of the same text. Here, we see that STVO and OST, together with STVIE, NIE and CIJA, form a distinct branch of similar items. This is because their distribution reflects an obvious semantic similarity: all these suffixes are used to derive abstract nouns, with different languages using them for different items and following different semantic models. While this observation may seem trivial in hindsight, it is significant that here it is arrived at based solely on corpus data, rather than secondary data, and extends also to resource-low languages such as Macedonian, in respect to which secondary sources may be difficult to come by.

Comparing the distribution of these morphemes in the corpus thus affords insight into the meanings of these suffixes. In general, many other items can be operationalized in a similar fashion, e.g. reflexive pronouns, pronominal forms, the use of tense, aspect, indefinite pronouns, prepositions, case forms or names; for representative case studies, see von Waldenfels (2014; 2015a). Translation is thus extremely valuable in providing a knowledge-richt type of annotation that links all the texts on a semantic level, making quick and rather comprehensive assessments of a wide range of issues possible. In general, our approach provides data-driven notions of relative semantic similarity, rather than semantic substance. Comparison of distribution provides a way to structure the data in non-arbitrary ways, rather like the semantic maps approach in linguistic typology (Haspelmath, 2003). More modes of analysis, such as the clustering of language-specific

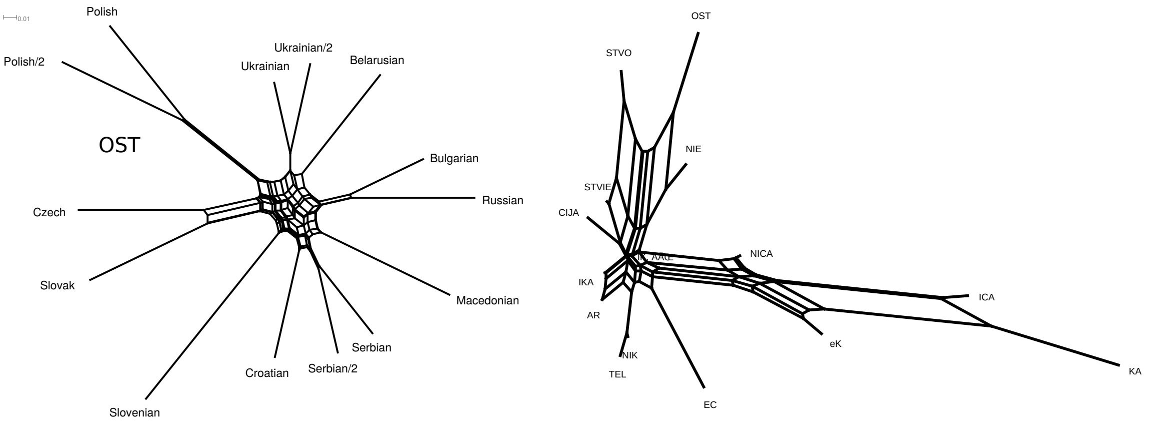


Figure 3: NeighborNets (Bryant and Moulton, 2004) represening similarity of use of nouns derived with the suffix class OST in different Slavic versions of the same text (left); similarity of suffix classes across Slavic (right).

members of these suffix classes, are not implemented yet; for these and more details on the method, see (von Waldenfels, 2015b).

ParaViz uses its own built-in parallel corpus - there is no possibility (at least at this stage) to use custom corpora. Currently the system uses ParaSol, a multilingual parallel corpus primarily geared towards linguistic contrastive and typological research[2] (von Waldenfels, 2011). ParaSol focuses on Slavic, but also includes Romance, Germanic, Finno-Ugric, Greek, Armenian and other languages. Most languages are lemmatized and POS-tagged; a subset of the corpus is word aligned using UPLUG (Tiedemann, 2003). We hope that with time, the system will be used by researchers interested in very different language combinations.

3 The application

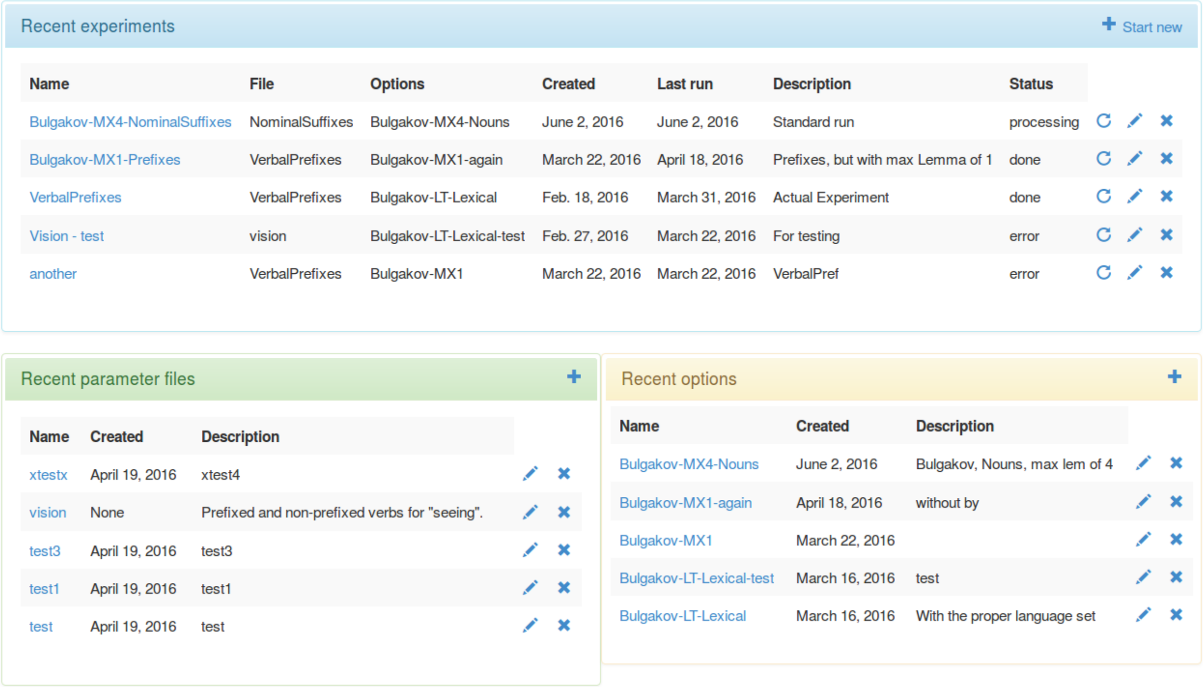
While ParaViz in its off-line version is a functional and powerful research instrument, it was developed in a patchwork-like way for a specific research project. It involves many manual steps and uses many technologies (including Perl, XSLT, Java, Python, R and specialized visualization software), running on a Linux operating system. This makes it virtually useless to anyone but its creators. The aim of the online version is thus to take care of all technicalities and provide users with an easy way of conducting their own research with parallel corpora.

ParaViz is developed with Django, a state-of-the-art Python powered web framework which is database oriented. The online version exhibits an extremely simple layout and has many performance issues, but the main functions are implemented and working properly. Generally, we do not plan to implement all functions of the offline system, but rather aim to provide users with all the relevant data files so that they can be used with other tools.

Registered users get assigned their own project space on the server where they may then create their own experiments. Experiment are the basic data objects: they can be created, run, examined, changed or deleted by the user. An experiment represents a linguistic problem that is investigated; in the above example, this would be the use of cognate noun suffixes in different Slavic languages. Each experiment consists of two input objects: a parameter file (see section two above) and a set of options that defines a configuration of texts, languages, and a number of parameters geared to managing lexical and selection effects during the experiment. At this stage parameter files have to be prepared and uploaded by the user; in the future, a graphical tool for the creation of such files may be added. Both parameters and options are saved, modified, and copied independently of experiments, for which they are reused.

The main page of the user interface (figure 4) is minimalistic, listing experiments, parameter

[2] http://www.parasolcorpus.org



Recent experiments

Name	File	Options	Created	Last run	Description	Status			
Bulgakov-MX4-NominalSuffixes	NominalSuffixes	Bulgakov-MX4-Nouns	June 2, 2016	June 2, 2016	Standard run	processing	↻	✎	✖
Bulgakov-MX1-Prefixes	VerbalPrefixes	Bulgakov-MX1-again	March 22, 2016	April 18, 2016	Prefixes, but with max Lemma of 1	done	↻	✎	✖
VerbalPrefixes	VerbalPrefixes	Bulgakov-LT-Lexical	Feb. 18, 2016	March 31, 2016	Actual Experiment	done	↻	✎	✖
Vision - test	vision	Bulgakov-LT-Lexical-test	Feb. 27, 2016	March 22, 2016	For testing	error	↻	✎	✖
another	VerbalPrefixes	Bulgakov-MX1	March 22, 2016	March 22, 2016	VerbalPref	error	↻	✎	✖

Recent parameter files

Name	Created	Description		
xtestx	April 19, 2016	xtest4	✎	✖
vision	None	Prefixed and non-prefixed verbs for "seeing".	✎	✖
test3	April 19, 2016	test3	✎	✖
test1	April 19, 2016	test1	✎	✖
test	April 19, 2016	test	✎	✖

Recent options

Name	Created	Description		
Bulgakov-MX4-Nouns	June 2, 2016	Bulgakov, Nouns, max lem of 4	✎	✖
Bulgakov-MX1-again	April 18, 2016	without by	✎	✖
Bulgakov-MX1	March 22, 2016		✎	✖
Bulgakov-LT-Lexical-test	March 16, 2016	test	✎	✖
Bulgakov-LT-Lexical	March 16, 2016	With the proper language set	✎	✖

Figure 4: Main user interface

files and option sets. Each list can be expanded to show all of its objects and each object can be viewed in detail or removed. Users can also create new experiments or option objects on the basis of existing ones.

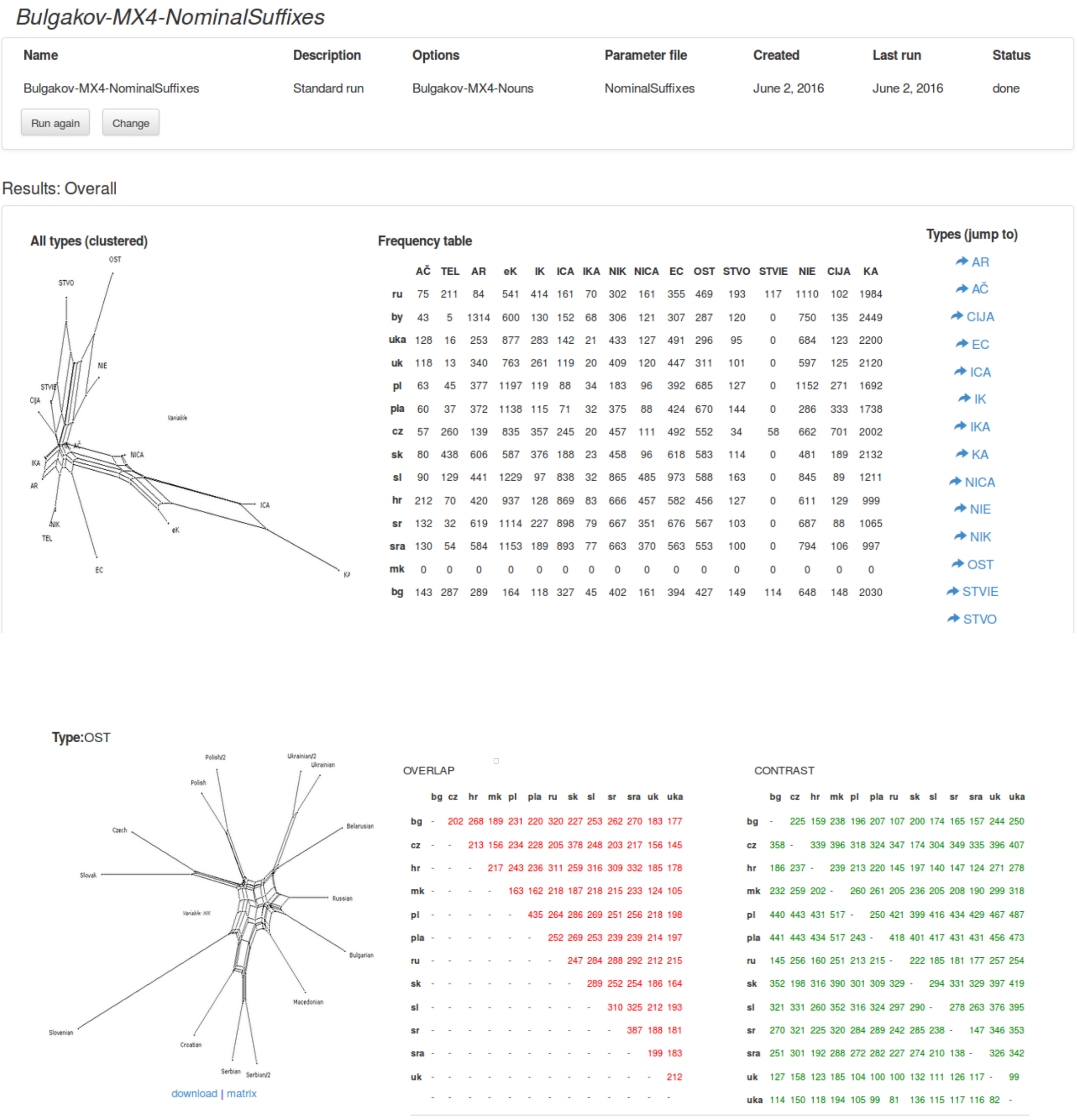
When users hit the "Run" button, the system checks if there are enough resources (both CPU and memory usage) and either starts to execute scripts or adds the experiment to the queue of experiments to be run as soon as resources are available. The experiment status changes from 'created' to 'waiting', 'processing', and finally, either 'done' or 'error'. The processing time is typically a few minutes, depending on the experimental settings.

After the experiment has been processed, the user is provided with an overview of its results as shown in figure 5. The system first gives a graph with the clustering of types (here: nominal suffixes) and their basic frequencies broken down by language and type. Then, for each type, it offers a graph of its use across languages, and a matrix of languages that provides the number of eqivalent word forms where two versions agree in using this type, and a second matrix with the number of equivalent word forms where only one of two versions use it - that is, of the data that is visualized in the graph to the left. Clicking on these numbers will open a new window with a random sample of color-coded corpus examples (see above figure 2) illustrating this case. Finally, files with classifications representing the use of the variables in the corpus, as well as the nexus files used to generate the networks, can be downloaded alongside their pdf versions for publication.

Given the right operationalization, thus, the tool provides users with both qualitative data allowing them to assess operationalization and the language data itself, as well as an aggregate perspective based on the same data, allowing them to proceed quickly in the analysis of the comparative issue they are engaged with.

4 Summary and outlook

In our paper, we have presented a tool that uses a semantically rich multilingual resource, translated texts, for the comparison of the use of multilingual categories. The tool was originally developed as an offline set of scripts and procedures developed for a specific project. Here, we aim to make it available to the research community at large in order to make the methodological



Bulgakov-MX4-NominalSuffixes

Name	Description	Options	Parameter file	Created	Last run	Status
Bulgakov-MX4-NominalSuffixes	Standard run	Bulgakov-MX4-Nouns	NominalSuffixes	June 2, 2016	June 2, 2016	done

Results: Overall

All types (clustered)

Frequency table

	AČ	TEL	AR	eK	IK	ICA	IKA	NIK	NICA	EC	OST	STVO	STVIE	NIE	CIJA	KA
ru	75	211	84	541	414	161	70	302	161	355	469	193	117	1110	102	1984
by	43	5	1314	600	130	152	68	306	121	307	287	120	0	750	135	2449
uka	128	16	253	877	283	142	21	433	127	491	296	95	0	684	123	2200
uk	118	13	340	763	261	119	20	409	120	447	311	101	0	597	125	2120
pl	63	45	377	1197	119	88	34	183	96	392	685	127	0	1152	271	1692
pla	60	37	372	1138	115	71	32	375	88	424	670	144	0	286	333	1738
cz	57	260	139	835	357	245	20	457	111	492	552	34	58	662	701	2002
sk	80	438	606	587	376	188	23	458	96	618	583	114	0	481	189	2132
sl	90	129	441	1229	97	838	32	865	485	973	588	163	0	845	89	1211
hr	212	70	420	937	128	869	83	666	457	582	456	127	0	611	129	999
sr	132	32	619	1114	227	898	79	667	351	676	567	103	0	687	88	1065
sra	130	54	584	1153	189	893	77	663	370	563	553	100	0	794	106	997
mk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bg	143	287	289	164	118	327	45	402	161	394	427	149	114	648	148	2030

Types (jump to)

- AR
- AČ
- CIJA
- EC
- ICA
- IK
- IKA
- KA
- NICA
- NIE
- NIK
- OST
- STVIE
- STVO

Type:OST

OVERLAP

	bg	cz	hr	mk	pl	pla	ru	sk	sl	sr	sra	uk	uka
bg	-	202	268	189	231	220	320	227	253	262	270	183	177
cz	-	-	213	156	234	228	205	378	248	203	217	156	145
hr	-	-	-	217	243	236	311	259	316	309	332	185	178
mk	-	-	-	-	163	162	218	187	218	215	233	124	105
pl	-	-	-	-	-	435	264	286	269	251	256	218	198
pla	-	-	-	-	-	-	252	269	253	239	239	214	197
ru	-	-	-	-	-	-	-	247	284	288	292	212	215
sk	-	-	-	-	-	-	-	-	289	252	254	186	164
sl	-	-	-	-	-	-	-	-	-	310	325	212	193
sr	-	-	-	-	-	-	-	-	-	-	387	188	181
sra	-	-	-	-	-	-	-	-	-	-	-	199	183
uk	-	-	-	-	-	-	-	-	-	-	-	-	212
	-	-	-	-	-	-	-	-	-	-	-	-	-

CONTRAST

	bg	cz	hr	mk	pl	pla	ru	sk	sl	sr	sra	uk	uka
bg	-	225	159	238	196	207	107	200	174	165	157	244	250
cz	358	-	339	396	318	324	347	174	304	349	335	396	407
hr	186	237	-	239	213	220	145	197	140	147	124	271	278
mk	232	259	202	-	260	261	205	236	205	208	190	299	318
pl	440	443	431	517	-	250	421	399	416	434	429	467	487
pla	441	443	434	517	243	-	418	401	417	431	431	456	473
ru	145	256	160	251	213	215	-	222	185	181	177	257	254
sk	352	198	316	390	301	309	329	-	294	331	329	397	419
sl	321	331	260	352	316	324	297	290	-	278	263	376	395
sr	270	321	225	320	284	289	242	285	238	-	147	346	353
sra	251	301	192	288	272	282	227	274	210	138	-	326	342
uk	127	158	123	185	104	100	100	132	111	126	117	-	99
uka	114	150	118	194	105	99	81	136	115	117	116	82	-

Figure 5: Result page: at the top, experiment description and overview of results with basic statistics. Below, one of a number of rows with per-type results providing NeighborNet graphs and matrices of corpus examples where translations into two languages do or do not agree in using the suffix in question. The matrices give the number of cases; clicking on the number will open a new window with a color-coded random sample of these cases in the corpus.

results of our research available to other scholars and open our results to replication, aims that have become, in our view, both more relevant and more readily attainable with the development of digital humanities.

The basic functions of this tool are grounded in a distributional model of semantics that utilizes translation as semantic annotation providing a data-driven method to derive comparative models of meaning of a large range of possible linguistic variables. While the original research was done only on Slavic languages, the tool is language independent and the corpus data it is used on involves many Romance, Germanic, Finno-Ugric and other languages.

Rather than building an offline version that would cater to a computationally literate community only, we have opted to prepare an online version built around the existing scripts. Focusing on functionality and transparency, we have devised a simple interface that enables the researcher to perform basic comparisons of a wide range of user-definable variables based on their use in the parallel corpus and download the relevant categorizations for further use. In the future, we plan to add a number of further analytic functions and, if time allows, provide a graphical tool for the construction of the parameters that form the basis of the experiments.

Acknowledgment

Funding by the Swiss National Science Foundation, grant 151230 *Convergence and divergence of Slavic from a usage based, parallel corpus driven perspective*, is gratefully acknowledged.

References

David Bryant and Vincent Moulton. 2004. Neighbor-net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, 21(2):255–265.

Michael Cysouw and Bernhard Wälchli, editors. 2007. *Parallel Texts: Using translational equivalents in linguistic typology. Special Issue of STUF 60/2.*

Östen Dahl. 2014. The perfect map: Investigating the cross-linguistic distribution of tame categories in a parallel corpus. In Benedikt Szmrecsanyi and Bernhard Wälchli, editors, *Aggregating Dialectology and Typology: Linguistic Variation in Text and Speech, within and across Languages*, pages 268–289. De Gruyter Mouton, Berlin, New York.

Martin Haspelmath. 2003. The geometry of grammatical meaning: semantic maps and cross-linguistic comparison. In M. Tomasello, editor, *The new psychology of language: Cognitive and functional approaches to language structure. Vol. 2*, pages 211–42. Laurence Erlbaum Associates, Mahwah, NJ.

Magnus Sahlgren. 2008. The distributional hypothesis. *Rivista di Linguistica*, 20(1):33–53.

Jörg Tiedemann. 2003. *Recycling Translations – Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing.* Ph.D. thesis, Uppsala University, Uppsala, Sweden. Anna Sågvall Hein, Åke Viberg (eds): Studia Linguistica Upsaliensia.

Ruprecht von Waldenfels. 2011. Recent developments in parasol: Breadth for depth and xslt based web concordancing with cwb. In Daniela Majchráková and Radovan Garabík, editors, *Natural Language Processing, Multilinguality. Proceedings of Slovko 2011, Modra, Slovakia, 20–21 October 2011*, pages 156–162, Bratislava. Tribun EU.

Ruprecht von Waldenfels. 2014. Explorations into variation across slavic: taking a bottom-up approach. In Benedikt Szmrecsanyi and Bernhard Wälchli, editors, *Aggregating Dialectology and Typology: Linguistic Variation in Text and Speech, within and across Languages*, pages 290–323. De Gruyter Mouton, Berlin, New York.

Ruprecht von Waldenfels. 2015a. Inner-slavic contact from a corpus driven perspective. In Emmerich Kelih, Stefan Michael Newerkla, and Jürgen Fuchsbauer, editors, *Lehnwörter im Slawischen: Empirische und crosslinguistische Perspektiven*, pages 237–263. Peter Lang, Frankfurt.

Ruprecht von Waldenfels. 2015b. The paraviz tool: Exploring cross-linguistic differences in functional domains based on a parallel corpus. In Gintarė Grigonytė, Simon Clematide, Andrius Utka, and Martin Volk, editors, *Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015, May 11–13, 2015, Vilnius, Lithuania.*

Author Index